MACHINE SPEECH, HUMAN SPEECH

MSc Thesis *(Afstudeerscriptie)*

written by

Raya Mezeklieva

under the supervision of **Dr Fausto Carcassi** and **Dr Giorgio Sbardolini**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
July 10, 2026	Dr Fausto Carcassi
	Dr Giorgio Sbardolini
	Dr Katrin Schulz
	Dr Sonia Ramotowska

Institute for Logic, Language and Computation

Abstract

As large language models (LLMs) increasingly populate everyday writing, certain words and stylistic habits have come to be perceived as telltale signs of machine authorship, reportedly prompting some writers to avoid them. This thesis investigates this oppositional pathway of LLM-driven language change, in which human writers adapt their language to not resemble LLM output but to be read as unmistakably human. Because LLMs themselves learn from human language which may include these oppositional changes, we hypothesized that successful differentiation strategies would not remain exclusively human for long, making this an iterative process of adaptation and counter-adaptation. To test how this may affect our linguistic behavior over time, we designed an online experiment repurposing the chain-generation structure of the Iterated Learning paradigm. Participants iteratively edited a news article to make it read more convincingly as human-written and unlike a simulated, state-of-the-art LLM that had itself incorporated the editing strategies of previous participants. We assessed (1) the sustainability of this process, based on participants' own judgments of whether their edits improved the text's signaling of human-authorship; (2) whether the strategies underlying participants' edits changed systematically across generations; and (3) whether such changes were reflected in changes in the lexical diversity, syntactic complexity, and orthographic and grammatical correctness of the texts themselves over time. We found that participants could keep finding means to differentiate text from machine-generated output across several iterations. Editing rationales showed no consistent linear change across generations, though exploratory analyses suggested a possible non-monotonic, reaction-and-counterreaction pattern along some dimensions. At the level of the texts, lexical diversity declined steadily across generations, while syntactic complexity showed no comparable systematic change and error rate only increased slightly in early generations. These findings provide some of the first experimental evidence that individuals can repeatedly adjust their linguistic behavior to differentiate it from machine-generated text, while also suggesting that, in the absence of mechanisms for coordinating such adaptations, this process may tend toward lexical simplification rather than the lexical innovation more typically associated with language use for identity signaling.

Acknowledgements

This work came to be thanks to several people who believed in me. I am especially grateful to my supervisors, Giorgio and Fausto, for enthusiastically and wholeheartedly engaging with the topic I proposed and for going out of their way to make it possible for me to conduct the study I envisioned. Thank you for giving me the space and resources to develop my own ideas but also for your thoughtful input and criticism. I want to thank Karolina as well for her caring and timely mentorship.

I also owe to the friends and family who participated in preliminary versions of the experiment for their invaluable feedback, as well as to everyone with whom I discussed my work over the past months. Thank you to Ralitsa and Miryana for inspiring me to find a topic I am genuinely passionate about. To my parents, thank you for the unconditional support during my studies abroad. In many ways, I have been fortunate to have had the resources and stability to wrestle primarily with my own doubts. And Dennis, thank you for always being patiently there.

Contents

Chapter 1

Introduction

In 2024, an abrupt and substantial shift in the use of English caught the attention of researchers and the wider public. Several words that had remained relatively stable in frequency for decades suddenly pervaded the language of scientific publications, podcasts, and academic talks (Yakura et al., 2025). Among the most infamous examples was the verb *delve*, otherwise unremarkable if slightly literary, whose frequency increased by more than 1,500% across major scholarly databases between 2022 and 2024 (Kousha & Thelwall, 2026).

The timing of these shifts left little room to question their cause. They neatly followed the public release and the widespread adoption of ChatGPT (Hu, 2023; OpenAI, 2022), including among researchers (Van Noorden & Perkel, 2023). This chatbot is powered by a large language model (LLM), that is, a model of language trained to generate text similar to human writing. The temporal alignment is unlikely to be coincidental as all surging words are also overrepresented in the output of the LLM powering ChatGPT compared to pre-2022 human writing (Yakura et al., 2025). As the same words have also crept into unscripted conversations and academic talks, growing in spoken use by roughly 25 to 50% a year following ChatGPT's release (Yakura et al., 2025), it appears that this influence cannot be attributed solely to LLM-based ghostwriting. Rather, repeated exposure to LLM output, whether through direct conversation with chatbots or passive consumption of LLM-assisted writing, seems to be gradually reshaping the vocabulary speakers reach for themselves (Yakura et al., 2025). Thus, in the span of two years, a handful of LLMs appear to have left a measurable mark on English language use by quietly permeating it.

Yet this story of passive, exposure-driven uptake is only half the picture. The same period that saw *delve* spread also saw it become a target of suspicion. Together with other words and linguistic features, it was flagged by researchers and online communities as a giveaway of LLM involvement (Chaib, 2026; Fink, 2025; Juzek & Ward, 2024; Shapira, 2024). Soon after, the frequency of its use measurably declined, even as other, less-discussed ChatGPT-favored words such as *significant* continued climbing (Geng & Trotta, 2025). The same retreat was reported anecdotally by individuals on social media, who described deliberately dropping words like *delve* (AskingAboutChatGPT, 2026) or the em dash (Ice_-Efficient, 2026; Si1Fei1, 2025) due to their becoming commonly associated with LLMs. This points to a second pathway of LLM-driven language change, whereby as certain words and stylistic habits become stigmatized as signatures of machine authorship, writers begin reworking their language to not resemble LLM output and to be unmistakably read as human.

While this second pathway has received far less attention than the first, it falls within an established sociolinguistic tradition of studying language as a tool for identity construction. Ample research has

established that language is not merely a vehicle for reference, but a malleable resource speakers draw on to construct identity, position themselves with respect to social groups, and distance themselves from those they do not wish to be associated with (Bucholtz & Hall, 2005; Eckert, 2019; Walters, 1987). Avoiding *delve* may thus not be so different from the many other ways speakers have shaped their language to signal who they are, and who they are not. What is new is the out-group: not a rival social class or community, but a machine.

The language of this machine out-group, however, differs from that of any human one in that it is derived from the human in-group's language. Both the data LLMs are trained on and the feedback used to align them originate from humans, so each stage of their training encodes human communicative practices, including attempts to distinguish one's writing from machine-generated text (Juzek & Ward, 2024; Yakura et al., 2025). As a result, differentiation strategies may not remain effective for long, as they may themselves enter the linguistic environment from which future models learn. The avoidance of *delve* or the em dash may therefore not be the end of the story, but the beginning of one of iterative mutual adaptation and counter-adaptation, in which human writers continuously adjust their language to differentiate it from that of LLMs, which in turn evolve following these linguistic adaptations.

This thesis sets out to investigate this second pathway of LLM-driven language change. Specifically, it aims to establish whether such linguistic differentiation could be sustained at all, rather than quickly exhausting the perceived means of signaling human authorship; whether the strategies writers draw on to do so evolve as the process unfolds; and whether those strategies leave detectable traces in the linguistic properties of the texts they produce.

To address these questions, we designed an online human-subjects experiment inspired from the Iterated Learning paradigm common in studies of language evolution (Kirby et al., 2008). It consisted of an iterated text editing task, in which every participant edits a text to make it read more convincingly as human-written and this edited text is then passed on to the next participant to edit following the same procedure, allowing successive adaptations to accumulate within a chain of edits.

We find that participants were able to sustain this iterative process across all planned iterations, with no chain exhausting its capacity to find further adaptations within the planned timeframe. The reasoning participants reported relying on, captured through free-text explanations of each edited sentence, did not shift in any consistent linear direction across generations, although exploratory analyses hinted at a non-linear pattern along some dimensions. The texts themselves, however, changed systematically with lexical diversity declining steadily across generations, while syntactic complexity and, to a lesser extent, orthographic and grammatical correctness remaining comparatively stable.

This asymmetry between a shifting lexicon and a stable syntax is, on its own, unsurprising. It aligns with the well-established tendency for lexical change to outpace syntactic change and for change generally to act first on more salient, consciously accessible features (Greenhill et al., 2017; Smitterberg, 2021). The lexical shift, however, is less orthodox. Rather than the relexicalization and lexical innovation typical of language varieties built around an in-group/out-group distinction, participants converged on simplification, narrowing the texts they produced rather than diversifying them, even as their explicit goal was to differentiate their writing from that of the machine.

This thesis thus has two sets of contributions. Theoretically, it extends accounts of language as a resource for identity construction and differentiation (Bucholtz & Hall, 2005; Eckert, 2019; Walters, 1987) to a setting in which the out-group is not a rival speech community but a generative model whose language is itself derived from, and continuously re-trained on, the in-group's own production and

feedback. In doing so, it identifies a candidate mechanism of language change, namely differentiation from a linguistically competent but non-human interlocutor, that does not map neatly onto existing accounts of contact-, or identity-driven change. Empirically, it moves the study of this phenomenon beyond the anecdotal reports and observational corpus data on which current evidence rests (Al-Sibai, 2025; Geng & Trotta, 2025; B. Hu, 2026), providing, to our knowledge, the first controlled experimental evidence that individuals can sustain such oppositional linguistic adaptation across iterations, and a first characterization of the strategies and linguistic changes through which they do so.

The remainder of the thesis is organized as follows. Chapter 2 situates the phenomenon within the literature on LLM-driven language change and on language as identity signaling, and closes by formally stating the three research questions and hypotheses. Chapter 3 details the design, materials, and analysis plan for the experiment summarized above. Chapter 4 reports the confirmatory and exploratory findings on sustainability, editing explanations, and text features in turn. Chapter 5 interprets these findings against the background literature, considers the limitations of the design, and outlines directions for future work. Finally, Chapter 6 summarizes the main contributions of the thesis and their implications for our understanding of language change in the age of LLMs.

Chapter 2

Background

This chapter develops the theoretical and empirical background for understanding large language models (LLMs) as agents of linguistic change and for investigating the forms of language change they may induce. It begins by introducing the relevant theoretical concepts in section 2.1, defining language and language change, surveying the internal, external, universal, and contingent factors that drive change, and situating language technologies, and LLMs in particular, among these influences. Given their ability to generate human-like yet still distinguishable language, LLMs are framed as producing a language variety we term "LLM-ese." The section ends by reviewing growing evidence that LLM-ese is permeating written and spoken English through a feedback loop of exposure and accommodation, establishing a first pathway of LLM-driven language change.

Section 2.2 then turns to a second, less explored pathway, arguing that the unreliability of human and automated authorship judgments, together with the social costs associated with being suspected of LLM use, creates incentives for writers to differentiate their language from LLM-ese. Drawing on sociolinguistic theories of identity signaling and anti-languages, this process is conceptualized as a form of oppositional identity signaling driven by strategic behavior rather than passive uptake. The section ends by recasting this process as an iterative arms race between co-evolving human writing practices and LLM output and by motivating the use of the Iterated Learning paradigm to study this dynamic empirically. The chapter closes by defining the research questions and hypotheses that guide the remainder of this work.

2.1 Large Language Models as Agents of Language Change

2.1.1 Notions of Language Change

Defining Language and Language Change

Several conceptualizations of language and language change coexist in the literature. Formalist and generative approaches view language as an innate, rule-governed system that reflects an innate language faculty or I-language (Chomsky, 1986). In this tradition, language change is understood as grammar change arising during language acquisition as learners restructure linguistic rules on the basis of the input they receive. Change is therefore conceived as discrete or abrupt and driven by internal pressures, such as the need to resolve ambiguity in the input.

Alternative accounts place greater emphasis on language as a social practice rather than an abstract mental system. Usage-based approaches argue that linguistic structure emerges from speakers' experience

with language in use and is shaped by processes such as repeated exposure strengthening linguistic patterns (entrenchment) and the extension of existing patterns to new forms (analogy), making frequency of use a central explanatory factor in language change (Bybee, 2015). Language itself may be understood as "the population of utterances in a speech community" (Croft, 2000, p. 26).

Consistent with the focus on language in use, variationist sociolinguistics, associated particularly with the work of Labov, treats linguistic variation as an inherent property of language and studies how socially conditioned differences in usage spread through speech communities over time (e.g., Labov, 1994). Language may thus be defined as "the instrument of communication used by a speech community, a commonly accepted system of associations between arbitrary forms and their meanings" (Labov, 1994, p. 9, cited in Smitterberg, 2021). Within this view, the language that varies across speakers, contexts, and time periods is more precisely described in terms of *varieties*, referring to the particular type of language singled out for description or comparison in sociolinguistic, dialectological, or corpus linguistic studies and defined by region, social group, register, or some other criterion (Grieve et al., 2025). The most granular variety is the idiolect, which refers to an individual's unique use of a language.

These social and usage-based perspectives converge with a third, more formal characterization of language as convention. Convention-based accounts view language as a system of socially coordinated regularities in action and belief, maintained and modified through repeated interaction among speakers (Lewis, 2008), making explicit the coordination problem that Labovian and usage-based notions of "shared" and "commonly accepted" language already presuppose. Thus, all three views conceive of language as emerging from repeated communicative interactions among language producers and being continuously shaped by patterns of use. This is the conceptualization we adopt in this thesis.

Such a construal of language allows conceptualizing language change as a process that unfolds through the linguistic behavior of individual speakers, following Croft (2000), who distinguishes two processes of change: innovation and propagation. Innovation refers to the introduction of a new feature into a speaker's idiolect, whereas propagation refers to the spread of that innovation through the speech community. Some scholars therefore treat propagation as a necessary condition for a linguistic innovation to qualify as a change in the language more broadly (Smitterberg, 2021). Notably, changes in the relative frequency of linguistic features reflect the presence of language change, as they indicate shifts in the patterns of use that characterize the language of a speech community (Smitterberg, 2021). For the purposes of this thesis, this definition is important because it allows language change to be traced through actual patterns of use in recorded language productions and because it accommodates socially motivated change, which is essential for understanding how LLMs may influence language both through passive uptake and through deliberate differentiation from LLM-associated forms.

Having established what counts as language and language change, we now turn to why and how such change happens. The following subsection reviews several factors that have been identified in the literature as driving language change, providing the theoretical background to reason about why LLMs may be expected to influence language and what forms of change they may induce.

Factors in Language Change

The literature distinguishes between multiple factors driving language change that operate at different levels of explanation. One distinction is drawn between language-internal and language-external factors, depending on whether the source of change lies in the structure and use of the linguistic system itself or in broader social and environmental conditions. Additionally, factors may be universal or contingent,

depending on whether their effects are observed across languages and historical contexts or arise from circumstances specific to a particular speech community and period. Language change is shaped by the interaction of various such factors, the most relevant of which for the present discussion are summarized below.

Language-internal changes arise from properties of the linguistic system itself and the cognitive processes operating on it, including pressures toward economy and ease of articulation or processing, analogical extension of existing patterns to new items, and reanalysis of ambiguous structures into new grammatical configurations (Bybee, 2015). Language-external factors, on the other hand, originate outside the system proper, in contact between speech communities, social stratification, or changes in the communicative environment (Aitchison, 2013; Bybee, 2015). These drive, for example, lexical innovations prompted by wars or new technologies, and the divergent trajectories of British and American English following their geographical and social separation (Bochkarev et al., 2014). The present discussion primarily concerns the latter category of factors. LLMs can be understood as a change in the communicative environment that exposes speakers and writers to a novel source of linguistic input, while the accommodation to or avoidance of LLM-generated language examined in this thesis arises from social motivations tied to how linguistic forms are evaluated and what identities they are taken to signal. Both mechanisms therefore operate through language-external factors rather than through pressures internal to the linguistic system itself.

Beyond the source of change, a second set of factors differ in the generality of their effects. Some patterns recur across languages and historical periods regardless of which speech community is involved, while others are tied to the particular sociohistorical circumstances of one community at one point in time (Bochkarev et al., 2014). Among the broadly recurrent tendencies, frequency is one of the most robust predictors of stability. More frequent words tend to have slower frequency shifts over time, with dominant lexical items remaining such as their repeated use reinforces their entrenchment in speakers' mental representations, making them less susceptible to displacement (Bybee, 2015; Pagel et al., 2019). The core lexicon identified in early lexicostatistical work is accordingly considered among the most stable parts of any vocabulary (Greenhill et al., 2017; Swadesh, 2006), whereas rarer words, including jargon and domain-specific terms, are more sensitive to historical events and social transformation (Bizzoni et al., 2020; Bochkarev et al., 2014; Hall et al., 2008).

Frequency interacts with another dimension that reinforces stability, namely the degree to which a form is tied to specific referents in the world. Content, open-class lexical items, such as nouns or verbs, name entities or actions and are therefore more directly exposed to shifts in the external environment. This is clearly illustrated, for example, in the evolution of scientific jargon over time in tandem with the relevant scientific discoveries (Bochkarev et al., 2014; Degaetano-Ortlieb & Teich, 2018). In contrast, function words, such as prepositions and conjunctions, and other grammatical elements serve logical and structural purposes rather than referential ones, expressing relations between propositions and maintaining the organizational regularities of the language (Bybee, 2015; Hopper, 2003). The closed-class status of these lexical items and the tight integration of these elements into grammatical paradigms shields them from such external pressures, and their usual high frequency compounds this, making them among the more diachronically conservative elements of a language (Degaetano-Ortlieb & Teich, 2018; Greenhill et al., 2017).

A related factor is the degree to which a feature is perceptually available to speakers. Even among the generally more slowly changing grammatical elements, those that are more overt (e.g., order-related)

are often more susceptible to change than abstract ones that operate beneath the threshold of ordinary attention or awareness (e.g., the relevance of grammatical categories to the language's morphosyntax such as the marking of future tense on the verb) (Greenhill et al., 2017). Salience can also arise from social evaluation, with features that are socially marked or indexically associated with particular groups or identities being more likely to be noticed and therefore more susceptible to change (Aitchison, 2013; Labov, 1994).

These factors do not operate in isolation, and the same change can be shaped jointly by constraints that hold regardless of community and by circumstances specific to one. Bochkarev et al. (2014) illustrate this through the partial separation of British and American English. These two varieties descended from a common source but were subject to distinct historical environments since their geographical divergence (e.g., different wars or social trends). They find that broad statistical tendencies, such as the relative stability of frequent words, hold across both varieties, while the specific trajectories of individual words diverge in ways traceable to each community's particular history.

In the context of the present work, LLMs represent a recent change in our communicative environment, which can be expected to influence language change in a way funneled by the above principles and factors. The following section examines this contingency in more detail in light of how language technology precedents have historically shaped language varieties and how LLMs align with and differ from these precedents.

Language Technology in Language Change

A sociohistorical contingency with particularly profound influence on language change is the adoption of technologies for language and cultural transmission (Vanmassenhove, 2025). As any change in the communicative environment, they would be expected to introduce new terminology related to the technologies themselves (Bochkarev et al., 2014; Degaetano-Ortlieb & Teich, 2018), but they also exert a more profound stylistic and structural influence. For example, the printing press has been associated with the standardization of vernacular languages in Europe (Sasaki, 2017), while the internet has been associated with the emergence of registers characterized by brevity and informality (Crystal, 2011; Tagliamonte, 2016).

LLMs are a more recent and especially consequential case because, unlike earlier language technologies, they do not merely store, mediate, or correct language, but generate it (Vanmassenhove, 2025). In this respect, LLMs introduce a new kind of contingent pressure, of the sort illustrated by Bochkarev et al. (2014) for British and American English. Rather than two human populations drifting apart under distinct historical circumstances, it is human writers as a whole who are now embedded in a new communicative environment that includes a prolific and emulative yet statistically idiosyncratic machine interlocutor whose output increasingly enters the same textual record from which language change is tracked and fueled. The following sections present this situation as well as evidence of its already perceptible influence on English language use in the form of passive uptake of the characteristics of LLM output.

2.1.2 LLM-ese: An Emergent Language Variety

How Large Language Models Produce Language

Large language models (LLMs) are the latest development in a long line of language modeling systems that use statistical methods to generate linguistic forms based on previously observed text (Z. Wang et al., 2025). At its core, this process involves encoding a probability distribution over linguistic tokens (e.g., words or subword units) by estimating the likelihood of tokens occurring given particular linguistic contexts from text corpora (Jurafsky & Martin, 2026), and then generating text by iteratively sampling likely continuations from this distribution (Jurafsky & Martin, 2026).

LLMs differ from earlier models in their transformer-based architecture (Vaswani et al., 2017) and the enlarged and diversified training data that they use (Z. Wang et al., 2025), which allow them to generate text that is fluent, contextually appropriate and often difficult to distinguish from human writing (Brown et al., 2020). Recent LLMs are further shaped by additional post-training stages beyond the initial language modeling objective. These include, for example, supervised fine-tuning (or continued pre-training) on curated examples to adapt the model to a specific domain or task (Gururangan et al., 2020), and reinforcement learning from human feedback (RLHF), which involves the adjustment of the model's behavior based on human preferences for certain outputs over others, so as to align it more closely with human values and expectations (Kaufmann et al., 2025).

Because each of these stages draws on its own pool of texts and annotators, the characteristics of a model's output may owe to the composition of its fine-tuning and feedback data as much or even more so than to the corpus on which it was originally pre-trained (Juzek & Ward, 2024). The resulting model thus reflects a layered aggregation of heterogeneous linguistic signals, which we will see in the next section, produces a language variety that is distinguishable from human language on a macro level, even if it is often indistinguishable on a micro level.

LLM-ese

Although LLM output is often indistinguishable from human writing at the level of individual texts, systematic differences emerge when texts are considered in aggregate. At the lexical level, LLM output shows overuse of particular words and a less diverse vocabulary than aggregated human writing (Juzek & Ward, 2025; Muñoz-Ortiz et al., 2024; Padmakumar & He, 2024). At the morphosyntactic level, it contains more auxiliaries, numbers (and sequences thereof, Zamaraeva et al. (2025)) and symbols, which are typical of more objective language (Muñoz-Ortiz et al., 2024), as well as more pronouns (Muñoz-Ortiz et al., 2024). At the level of style and tone, it tends to be more uniformly positive and less affectively negative than human language (Khan et al., 2025; Muñoz-Ortiz et al., 2024). More generally, LLMs favor frequent and broadly applicable constructions, whereas humans more often use more idiosyncratic ones (Zamaraeva et al., 2025).

These recurring patterns motivate their casting as a new variety of English, which we will refer to as "LLM-ese"–a somewhat hygienized probabilistic merger of the writing styles of individuals whose idiolects are involved in some stage of LLMs' training process. This variety is coming in contact with the "human" one as people interact with LLM chatbots or consume LLM-generated content, leading to a propagation of LLM-ese features in human language. The next section will review evidence of this propagation already taking place in English and describe the mechanism through which it happens.

2.1.3 One Pathway of LLM-driven Language Change: The LLM-ese Feedback Loop

Despite their very recent popularization, LLMs appear to have already exerted a tangible influence on English language use. Both social media commentary (Jeremy Nguyen [@JeremyNguyenPhD], 2024; Shapira, 2024) and empirical research (Galpin et al., 2025; Geng & Trotta, 2025; Geng et al., 2025; Kobak et al., 2025; Kousha & Thelwall, 2026; Yakura et al., 2025) have pointed out a surge in the frequency of use of certain words, such as *delve* soon after the public release of ChatGPT in late 2022. Specifically, *delve* increased in use by over 1,500% across major scholarly databases between 2022 and 2024, while *underscore* (+1,000%), *intricate* (+700%), and *meticulous* (+600%) showed similarly dramatic growth (Kousha & Thelwall, 2026). In biomedical abstracts alone, *showcasing* increased by a factor of 10.7, *underscores* by 13.8, and *delves* by 28.0 compared to expected frequencies based on pre-ChatGPT trends (Kobak et al., 2025). This LLM-driven shift exceeds the impact of major world events; the frequency gap for LLM-related excess words in 2024 was at least twice as large as the peak of COVID-related vocabulary in 2021 (Kobak et al., 2025). What is more, the co-occurrence of LLM-preferred terms has also intensified dramatically: in 2024, 59.3% of papers mentioning *delve* also mentioned *underscore*, compared to only 1.3% in 2022, with correlation coefficients rising from near-zero (0.018–0.032) up to moderate (0.311–0.449) (Kousha & Thelwall, 2026; Schober et al., 2018).

While timing alone, on which these observations rely, cannot formally establish LLMs as the cause of the observed shifts, other lines of evidence also support this attribution. For example, all of the lexical items whose frequency has dramatically increased are among those overused by LLMs, that is, they are generated more frequently by LLMs than they were used in text written pre-ChatGPT. To determine this, Liang et al. (2024) and Yakura et al. (2025) compiled parallel corpora of human-written documents from the pre-ChatGPT era (2019–2022), drawn from sources including arXiv abstracts, bioRxiv preprints, Nature articles, Enron emails, student essays, news descriptions, and Wikipedia entries. These human texts are then edited by various GPT-family models (`GPT-3.5-turbo`, `GPT-4`, `GPT-4-turbo`, and `GPT-4o`) using standardized prompts (e.g., "proofread for grammatical accuracy with minimal distortion") to create matched LLM-edited corpora. Comparing word frequencies between each original document and its GPT-edited counterpart then isolates the lexical substitutions introduced by the model under conditions that hold content and register constant, yielding a directly attributable inventory of words systematically favored by GPT-family models over the human originals. It is this inventory that the studies draw on to track the uptake of GPT-favored vocabulary in genuinely human-authored production, whether published writing or unscripted speech.

Additionally, the increased use of these words is more pronounced in scientific fields whose practitioners are more familiar with and more likely to use LLMs, such as computer science and artificial intelligence, compared to fields that are less likely to use LLMs, such as humanities, social sciences, and mathematics (Kobak et al., 2025; Kousha & Thelwall, 2026; Liang et al., 2024). Causal inference methods further substantiate the role of LLMs in these shifts. For example, Yakura et al. (2025) used the synthetic control method from econometrics (Abadie et al., 2010) to compare the observed changes in the frequency of use of certain words in academic abstracts to a counterfactual scenario where LLMs were not released. Target words were those overused by GPT models (`GPT3.5-turbo`, `GPT-4`, `GPT-4-turbo`, `GPT-4o`), and the counterfactual scenarios for these "treated" words were constructed by weighting synonyms of the target words that weren't overused by the models. For example, *examine* and *explore* could be such synonyms for *delve*. They found that the observed changes were significantly

greater than what would be expected in the counterfactual LLM-free scenario.

Beyond increases in the frequency of individual lexical items, Galpin et al. (2025) found that entire synonym clusters were affected. For example, for focal words like *boast*, *crucial*, and *pivotal*, most or all words within their semantic groups increased in tandem rather than one word displacing its semantic neighbors. The clustering effect spans several part-of-speech categories, with changes concentrated around LLM-preferred words, which tend to be verbs, adjectives and adverbs, rather than content nouns. This indicates semantic and pragmatic shifts consisting in the use of more qualifications.

Changes in syntactic complexity, cohesion, and readability have also been observed following ChatGPT's release, as measured by mean clause length, mean sentence length, and T-units per sentence, indicating that LLMs contribute to a streamlining of sentence structures (Bao et al., 2025). At the same time, cohesion was found to decline with the proportion of linking words and lexical overlap between adjacent sentences decreasing significantly. Paradoxically, despite simplified syntax, readability was found to decrease with a 5.6% decline in the Flesch Reading Ease score and a 1.2% increase in the New Dale-Chall score from 2022 to 2023. This is explained by the increased use of specialized terminology and new concepts introduced by LLMs, combined with reduced explicit signaling of logical connections between propositions, which makes the text more difficult to comprehend.

In sum, there is ample evidence of LLM influence on the written English record, and this influence extends beyond the mere dissemination of LLM-related terminology to affect the frequency of use of certain topically unrelated words, as well as syntactic complexity, cohesion, and readability. However, the presence of these changes in the text produced by the scientific community does not necessarily indicate that they are due to the influence of LLMs on people's genuine language competence. Rather, they could be explained by the presence of LLM-generated or LLM-assisted writing in published content, which would introduce LLM-style vocabulary directly into the scholarly record without entailing deeper cognitive or linguistic change. The next section will review evidence that LLMs are indeed influencing people's language competence and not just the content of their writing, suggesting a real LLM-driven language change in progress.

LLM-Contaminated Text or Genuine Adoption of LLM-ese

Corpus-based linguistics has traditionally treated written and spoken productions as reflecting the language of a community of speakers (e.g., Degaetano-Ortlieb & Teich, 2018). This assumption is weakened in post-2022 corpora as LLMs are increasingly used to generate or assist in the production of written content. A text retrieved from a corpus may no longer reflect its named author's own linguistic behavior, but rather, in part or in whole, that of the model that helped produce it. Scientific publications are not exception. A 2023 *Nature* survey of 1,600 researchers revealed that almost 30% use generative AI tools for manuscript writing (Van Noorden & Perkel, 2023). Similarly, a 2024 Elsevier survey of 2,284 researchers reported that 31% of them have used AI for work-related purposes. This means that the shifts in language use observed in scientific publications could be simply due to the presence of LLM-generated content in the text rather than an actual change in language use by human authors (Goodchild et al., 2024).

Data from spontaneous speech, however, is not subject to this confound, as it cannot be generated or modified by LLMs. Yakura et al. (2025) used this idea to determine the extent to which the LLM-driven shifts in writing have propagated to genuine language use. They analyzed the frequency of ChatGPT words, including *delve*, *boast*, and *swift* 740,000 hours of podcast episodes and academic talks on

Youtube. They found that these words increased significantly in usage after the release of ChatGPT, with annual growth rates of approximately 25% to 50% compared to synthetic control words matched for pre-release frequency patterns (Yakura et al., 2025). The effect was statistically significant ($p < 0.05$) in YouTube academic talks and in conversational podcast episodes across Science and Technology, Business, and Education domains, though absent in Sports and Religion and Spirituality (Yakura et al., 2025). Similarly, Geng et al. (2025) found that words such as *significant*, *crucial*, *effectively*, and *comprehensive* substantially increased use in both written abstracts and oral presentations from major machine learning conferences (ICLR, ICML, NeurIPS) through 2024, suggesting an implicit, growing impact on human expression.

Having established the presence of LLM-driven shifts in genuine human language use, the next question is through what mechanism these shifts occur. The following section will describe how LLM-ese silently spreads through human speech via linguistic accommodation, and how this process is amplified by a feedback loop between both humans' and LLMs' exposure to language.

Mechanism: Linguistic Accommodation and Human-LLM Coevolution

The contamination of published text with LLM-generated content and the adoption of LLM-ese in speakers' genuine language use are mutually reinforcing processes that create a self-perpetuating feedback loop that amplifies their influence on language change. Repeated exposure to LLM-ese, whether through direct interaction with chatbots or passive consumption of LLM-generated content in published text, can lead to the internalization of these linguistic patterns. Through processes of linguistic accommodation and entrainment, speakers who regularly engage with AI-mediated communication progressively align their vocabulary, syntactic structures, and stylistic choices with those of the artificial interlocutors (Hancock et al., 2020; Székely et al., 2025). Importantly, this effect extends beyond direct users: as accommodated speech and writing spreads through social networks and professional discourse, even individuals who never interact directly with LLMs may adopt these patterns, leading to rapid and widespread shifts in language use at the societal level.

This bidirectional exchange initiates a "closed cultural feedback loop" in which cultural and linguistic traits circulate between humans and machines (Yakura et al., 2025). Machines originally trained with human input exhibit distinctive linguistic characteristics we called LLM-ese, which they then disseminate to millions of users. As humans adopt these traits in their own speech and writing, this LLM-influenced language feeds back into the training data for future models, creating a recursive cycle (Pedreschi et al., 2025). This coevolutionary dynamic between humans and LLMs has been portrayed as a threat of progressive erosion of linguistic and cultural diversity, as the feedback loop favors certain linguistic patterns over others, specifically the most statistically promiment, potentially leading to a form of "cultural singularity" where machine and human culture become increasingly indistinguishable (Yakura et al., 2025).

The story does not end here, however. As LLMs become more widely used and their linguistic patterns more recognizable, certain linguistic features start appearing "fake," (Sam Altman [@sama], 2025), being stigmatized as "AI tropes" (Chaib, 2026) or "AI telltale signs" (Fink, 2025), and consequently avoided by speakers who are aware of their association with LLMs (Harris, 2026; B. Hu, 2026). This suggests that the propagation of LLM-ese through passive accommodation and feedback is accompanied by a second, oppositional process of self-differentiation from LLMs that may exert its own effect on language. The remainder of this chapter develops this idea as a form of language change driven by

identity signaling against a machine interlocutor that successfully imitates human language.

2.2 A Second Pathway of LLM-driven Language Change: Oppositional Identity Signaling

In his seminal paper "Computing Machinery and Intelligence," Alan Turing proposed addressing the abstract philosophical question "Can machines think?" by recasting it into a practical test he called the Imitation Game (Turing, 1950). In the game, an interrogator communicates blindly with a human or with a machine via typed messages based on which the interrogator must determine the interlocutors' identities. Turing then asks whether the interrogator would fare no better chance than if the game were played with a man and a woman, that is, whether a machine can sustain a text-based exchange practically indistinguishable from a human's. Albeit two decades later than Turing anticipated[1], this question has found a positive answer in the name of LLMs (Biever, 2023; Chakraborty et al., 2023) as will be further shown below. However, its behaviorist framing, where what matters is not what is happening inside the system, but whether its outputs are indistinguishable from those of a human, has been criticized for failing to exhaust the concept of intelligence as present in the human mind (Block, 1990). This criticism is relevant here as it captures why authorship authenticity is a question that continues to be cared about by institutions and ordinary readers alike.

Block (1990)'s distinction between intelligence and intentionality is relevant in this sense. Intelligence, he argues, is future-oriented, that is, it concerns what a system has the capacity to do, whereas intentionality is past-oriented, requiring a causal history that makes a system's states genuinely represent the world. A system can satisfy the behavioral demands of intelligence while failing the historical requirement of intentionality. This point converges with Searle (1990)'s Chinese Room Argument: a system that manipulates symbols according to formal rules can produce appropriate outputs without understanding their content, and Block, despite rejecting Searle's broader conclusion, concedes that such a system may lack phenomenological consciousness and therefore genuine intentionality.

LLMs are intelligent in Block's sense, as they demonstrably produce fluent, contextually appropriate text, but whether their outputs are backed by anything resembling intentionality remains at the very least an open question. This openness, beyond a philosophical curiosity, has real-world implications for how we understand and value authorship. Authorship is not just a label for a piece of text. It comes with implications of intention, effort, accountability, and a human identity that can be evaluated, trusted, or blamed.

The gap between behavioral indistinguishability and attributed authorship is what generates the authenticity problem this section addresses. When authorship cannot be attributed reliably either by human judges or by automated detectors, but the answer nonetheless carries institutional and social consequences, the question becomes: can humans reliably signal human authorship through written language use, and if so, how? The test is now a differentiation game in which human writers must use language strategically to differentiate themselves from LLMs and signal their human authorship. Understanding this pressure, the conditions under which it arises, and the sociolinguistic mechanisms through which it can translate into language change is what this this section sets out to do.

[1]Earlier systems, such as the psychotherapist-imitating chatbot ELIZA have reaked some success, but in limited settings (Block, 1990; Weizenbaum & View Profile, 1983)

2.2.1 The Authenticity Problem

With their easy accessibility through chatbot interfaces and their ability to generate high-quality text, hardly distinguishable from human writing (Boutadjine et al., 2025; Weber-Wulff et al., 2023), LLMs have become widely used as writing aids, including by students (Freeman, 2025) and researchers (Goodchild et al., 2024; Van Noorden & Perkel, 2023). As a result, texts produced in recent years may invite suspicion of partial or complete LLM origin. Even though such judgments cannot be made reliably, they may nevertheless carry institutional and social costs. This can create an incentive for writers to manipulate the linguistic appearance of their text to signal human authorship and evade the negative social meanings associated with suspected LLM use.

The Failure of Human Judgment

Empirical studies pre-dating the public release of ChatGPT (then powered by `GPT-3.5`) showed that lay readers have great difficulty telling LLM text from human writing. Clark et al. (2021) had non-experts judge short news, recipe, and story excerpts generated by `GPT-3` or humans, and found that untrained participants distinguished them only at or near random-chance levels (even after basic training on examples of machine output, accuracy plateaued around 55% at best). Similarly, Kreps et al. (2020) found that people were unable to reliably distinguish `GPT-2`-generated news articles from human-written ones and often rated the former as equally or nearly as credible as the latter. In the literary domain, Köbis and Mossink (2021) found that participants could not reliably detect GPT-2-generated poems when the best samples were selected by a human, even when they knew that some of the poems were machine-written. While the poor detection abilities demonstrated in these early studies could be attributable to people's unfamiliarity with the possibility of machines generating fluent text, more recent studies corroborate these findings. Boutadjine et al. (2025) report that internet users achieve only 51–52% accuracy in classifying human-written and `GPT-3.5`-generated texts.

Follow-up studies looked into the reasons behind these failures and highlighted the role of miscalibrated heuristics. Jakesch et al. (2023) demonstrate that when evaluating self-presentations in professional, hospitality, and dating contexts, participants performed around chance at detecting LLM-authored texts, and their judgments were guided by intuitively plausible but empirically wrong cues. For example, readers tend to treat first-person pronouns, mentions of family and personal history as positive evidence of human authorship, even though current LLMs deploy precisely these features with high fluency. Conversely, they found that people over-associate grammatical errors or awkwardness with LLMs, despite modern systems producing text that is often more grammatically polished than human writing, although more recent studies report the opposite, i.e., people taking the lack of errors as a sign of LLM-generated text (Fiedler & Döpke, 2025; Georgiou, 2026; Russell et al., 2025). As a result, LLMs are perceived as "more human than human" (Jakesch et al., 2023, p. 4). This is further corroborated by Rathi et al. (2024) by showing that `GPT-4` was rated as more human-like than human-written texts. The most important predictor of these ratings was the presence of linguistic features that are typically associated with human writing, such as the use of first-person pronouns and mentions of family and personal history.

The picture slightly changes when considering experienced LLM users. Russell et al. (2025) find that people who frequently use LLMs for writing tasks achieve a near-perfect true positive rate and a minimal false positive rate when detecting text generated by `GPT-4o`, `Claude`, and `GPT-o1`. The majority vote among five such expert annotators correctly identified authorship in 299 out of 300 news

articles. These experts relied on subtle cues including specific "AI vocabulary" (e.g., *vibrant*, *crucial*, *significantly*), formulaic sentence structures, and assessments of originality. However, the study relied on a very small sample of nine evaluators, of which five experts, limiting the generalizability of the findings.

Similarly, domain expertise seems to help detection. Fiedler and Döpke (2025) find that among 63 professors evaluating excerpts from German theses, prior teaching experience slightly improved detection accuracy, although the overall performance remained only slightly above chance (57% accuracy for LLM-generated text and 64% for human texts), with no statistically significant difference between human and machine detection rates. Notably, professional-level LLM-generated texts proved most difficult to identify, with less than 20% of respondents classifying them correctly. On the other hand, Doru et al. (2025) report that medical and humanities experts achieved 70% accuracy in identifying ChatGPT-generated medical student essays, with their decisions based primarily on stylistic features, in particular redundancy, repetition, and thread or coherence issues, rather than content errors. The accuracy was independent of experts' content familiarity, suggesting that training readers to attend to specific linguistic attributes may enhance detection capabilities.

This hypothesis is corroborated by Milička et al. (2025) showing that participants receiving immediate feedback during a detection task achieved an average success rate of 65.1%, compared to 55.4% for those without feedback, which was a statistically significant improvement that also led to better confidence calibration. Participants without feedback made the most errors precisely when feeling most confident, whereas the feedback group learned to correct their misconceptions about LLM stylistic features.

Taken together, the evidence reviewed above suggests that authorship attribution has become an uncertain task. While some individuals can learn to identify characteristic features of contemporary LLM output, most people struggle to distinguish machine-generated from human-written text, and even successful judgments remain probabilistic rather than definitive. The result is a communicative environment in which suspicions of LLM use are common but difficult to substantiate. One possible response is to delegate the task to automated detection systems. The next section will show, however, that these systems are also unreliable, and that the uncertainty surrounding authorship is unlikely to be resolved by technological means.

The Failure of Automatic Certification

In response to growing concerns about jeopardized informational integrity posed by the hardly identifiable LLM-generated text, automated detection tools have proliferated, yet their reliability remains contested (Cabanac et al., 2021; Cheng et al., 2025; Fajt & Schiller, 2025; Tsigaris & Teixeira da Silva, 2026; Weber-Wulff et al., 2023). Empirical benchmarks show accuracy ranging approximately 50–90% across tools, including non-negligible false positives (falsely labeling human-written text as LLM-generated) and false negatives (failures to identify LLM-generated text) (Pudasaini et al., 2025).

Fully reliable detection, however, may be unatainable. While the architectures of the detection systems vary (Fariello et al., 2025), most [2] rely on distributional tendencies in the features of LLM-generated and human-written text (Tsigaris & Teixeira da Silva, 2026)—what we called "LLM-ese."

[2]Alternative approaches are emerging, but are also not infallible. For example, watermarking embeds detectable signals during text generation itself, but these can be removed through rephrasing (Kirchenbauer et al., 2024) and requires the cooperation of LLM developers (Májovský et al., 2024). Logging and information retrieval methods are more resistant to evasion techniques (Krishna et al., 2023), but require storing a model's output history, which poses privacy concerns, and also require cooperation with LLM developers (Májovský et al., 2024).

The exact nature of these features varies across tools, including statistical measures like perplexity and burstiness (Hans et al., 2024), stylometric features such as sentence length, use of certain words and phrases, syntactic complexity, and cohesion (e.g., StyloAI; Opara, 2025) and latent features extracted by machine learning classifiers (e.g., Verma et al., 2024). However, because these are population-level statistical tendencies from limited corpora, individual human-written texts that exhibit such features risk being falsely flagged as LLM-generated (false positives), while LLM-generated texts that steer away from these patterns may be missed (false negatives), especially when the texts are short (Kashtan & Kipnis, 2024; Masrour et al., 2025) and with mixed authorship (Artemova et al., 2025).

Májovský et al. (2024) further argue that detection efforts are "fundamentally flawed" (p. 1) because any improvement in discriminator performance can be directly repurposed to train better generators. This instantiates the dynamic of generative adversarial networks (GANs) whereby the discriminator's loss can be used as the generator's training signal: if a discriminator can assign a score indicating how "human-like" a text appears, that score can be optimized against by a generator so as to fool the discriminator (Májovský et al., 2024). Even if the discriminator temporarily gains the upper hand, a stronger generator can always be developed to render it obsolete. Moreover, evasion techniques, such as paraphrasing, word substitution, deliberate misspellings, punctuation changes, and Unicode homoglyph substitutions (Cai & Cui, 2023; Creo & Pudasaini, 2025; Krishna et al., 2023) can be applied post-hoc, independently of the generator. David and Gervais (2025) demonstrate this generator-detector arms race dynamic empirically by showing that a simple paraphrasing strategy can have an attack success rate of 78.6%–96.2% on state-of-the-art detectors, and training a new generator on the discriminator's feedback can reduce detection rates to near zero. Therefore, even if some tools show improvements in robustness (Masrour et al., 2025), the endeavor to ensure reliable detection may be vain (Májovský et al., 2024).

In the last two sections, we established that both human judgment and automatic detection tools fail as trustworthy arbitrators of text authorship. The next section will show that perceived LLM use, nonetheless, carries normative and social costs, which create incentives for writers to signal authentic authorship through their very language use.

The Cost of Suspected LLM Use

Suspicion of LLM use carries institutional and communicative repercussions. In academia, concerns about ghostwriting are becoming especially acute since authorship is closely tied to expertise, responsibility, and ethical standards (Padillah, 2024). Even when accusations are later withdrawn, individuals suspected of using LLMs may experience lasting reputational harm and psychologically corrosive presumptions of guilt (Liang et al., 2023; Showemimo, 2025). Similarly, in journalism, perceived machine authorship can reduce perceived source and message credibility (Jia et al., 2024; S. Wang & Huang, 2024). In workplace settings, managers who use AI for routine communications are perceived as less trustworthy when that assistance is suspected or disclosed (Cardon & Coman, 2025). Comparable effects have been observed more generally in AI-mediated communication, where perceived AI assistance in Airbnb host profiles is found to reduce trustworthiness judgments (Jakesch et al., 2019), and the belief that a communicator relies on smart replies can make them seem less cooperative, less affiliative, and more dominant (Hohenstein et al., 2023). In marketing, suspected AI involvement can also reduce positive word of mouth and customer loyalty, partly through authenticity concerns and moral disgust (Kirk & Givi, 2025). More generally, perceived AI authorship can lower perceived source

credibility, anthropomorphism, and intelligence (Henestrosa & Kimmerle, 2024).

These effects are not limited to judgments of the writer, but extend to judgments of the text itself. Content believed to involve AI is often evaluated less favorably even when it is identical to human-written content (Darda et al., 2022). In a large experimental study, participants rated creative writing as less enjoyable, less creative, lower in quality, and less appealing to audiences when they believed it was LLM-generated or LLM-assisted, relative to the same text presented as human-written; this devaluation was mediated by perceived authenticity and persistence (Raj et al., 2026).

In sum, the suspicion, whether founded or not, that a text was partially or fully generated by an LLM can lead to real communicative costs related to authenticity and trustworthiness. This creates a strong incentive for writers to manage the linguistic appearance of their text, not only to communicate content effectively, but also to signal human authorship and avoid the negative social meanings associated with LLM use. The next section will show how this linguistic signaling can constitute a second pathway of LLM-driven language change, one that is reactive and oppositional to the first pathway of passive accommodation and feedback.

2.2.2 Anti-LLM-ese as an Oppositional Identity-Building Strategy

Language as a Tool for Identity Signaling and Negotiation

As we have seen, living languages evolve to preserve their referential function in the changing environments of their speakers (Bochkarev et al., 2014). For example, major historical events such as wars and technological development are accompanied by the introduction of new lexical items in scientific discourse (Degaetano-Ortlieb & Teich, 2018).

At the same time, languages simultaneously evolve to fulfill the indexical and performative functions of signaling and negotiating social meanings (Eckert, 2012, 2019). Whereas reference concerns truth-conditional meaning, that is, the denotation of entities, qualities, and relations in the world, indexicality is the mechanism by which linguistic forms "point to" aspects of the social or personal context of utterance (Eckert, 2019). For example, morphosyntactic variables like negative concord or nonstandard *were* index institutional orientation and class-based stances, phonological variables like the backing of /æ/ in California English or postvocalic /r/ in New York City index regional identity, and prosodic and voice quality features (e.g., creaky voice, falsetto, whispery phonation) index affective states, emotional engagement, i.e., more "interior" qualities of personhood (Eckert, 2012). Similarly to phonological features, patterns of word use in text features also signal characteristics of the author's personality (Ireland & Mehl, 2014). What distinguishes such indexical sociolinguistic variables from symbolic reference is that they gain their meaning through their form and social source rather than through their denotation (Eckert, 2019). Notably, a linguistic form can directly index a contextual parameter (e.g., a southern accent indicating geographical origin), but it can also index a social meaning indirectly through its association with a particular social group (e.g., the same southern accent indexing a working-class identity) (Eckert, 2019, p. 753).

Whereas indexicality describes the mechanism by which linguistic forms accrue social meaning, performance describes the active, agentive deployment of these indexical resources to construct social identity (Eckert, 2019). Notably, performing identity through language (e.g., interjecting "Damn!") can be more "economical, more strategic, more believable, and more deniably intentional," that is, fulfilling their intended function while not overtly expressing it, than explicit assertions (e.g., "I'm angry") (Eckert, 2019, p. 752). Speakers build fluid, situated self-presentations, called personae, through

the strategic recombination of available indexical resources. Bucholtz and Hall (2005) emphasize this dynamicity of identity by construing it as a more or less conscious outcome of an interaction rather than a reflection of pre-existing identities. Identities are thus the product of a negotiation and rely on the interlocutors' perceptions and interpretations of the linguistic cues deployed in the interaction. It is this agentive, evaluative dimension of language use that distinguishes identity-driven change from change driven by passive exposure, as speakers do not merely absorb the forms they encounter, but select among them in ways that reflect and reinforce their social positioning.

The performative, identity-driven use of language connects back to the usage-based conception of change introduced in Section 2.1. Under Croft's distinction between innovation and propagation (Croft, 2000), a single act of identity is, in the first instance, just an innovation in one speaker's idiolect. What converts a recurring performative choice into language change proper is propagation, the process by which other speakers, recognizing and valuing the same social meaning, take up the same linguistic resource in their own performances. When enough speakers converge on using a feature to enact the same persona or stance, the feature's social meaning becomes conventionalized, or enregistered (Johnstone, 2016), as an index of that persona, at which point its use can spread independently of any one speaker's original communicative intent. When speakers perform identity through strategic linguistic choices in this way, these performances function as innovations that may propagate through the speech community, leading to language change.

Kristiansen (2014) distinguishes two forces driving this process: *density*, defined as "mechanical and inevitable effects of frequency" and based on Bloomfield's principle of density (Bloomfield, 1984), and *subjectivity*, defined as "motivation in terms of social evaluation and attitudes" and consistent with sociolinguistic perspectives on language change driven by social identity (Labov, 1994). Density operates mechanically: the more frequently a speaker is exposed to a linguistic form, the more likely that form is to become entrenched and reproduced, independently of any attitude the speaker holds toward it or toward its source. Subjectivity, by contrast, operates through evaluation: a form spreads, or is avoided, because of the social meaning speakers attach to it and to the group or persona it is taken to index. The LLM-ese feedback loop introduced in section 2.1 is a case of density at work, since the uptake of words like *delve* through repeated exposure to LLM output requires no attitude toward LLMs themselves, only sufficient contact with their output. The phenomenon we describe here is, by contrast, a case of subjectivity, where forms are avoided if they have come to index an unwanted social meaning, namely LLM-assisted writing. The limit case of such oppositional identity signaling is the anti-language, which we will discuss in the next section and which can be seen as a useful analogue for the phenomenon investigated in this thesis.

Anti-Languages as a Case of Oppositional Identity Signaling

The concept of anti-languages was introduced in Halliday (1976) to describe varieties of language used by marginalized groups to create a sense of identity and solidarity among their members, forming an anti-society in opposition to the mainstream culture.

A consequence of this oppositional orientation is that anti-languages are inherently unstable. Because their social function depends on difference from the mainstream, they must be continually renewed as forms become familiar, conventionalized, or absorbed into broader usage (Halliday, 1976). Anti-language is thus an iterative process of reconstitution, surviving only insofar as it can sustain the alternative social world it helps enact. It is a productive process featuring phonological, morphological,

and lexical innovation, the latter including relexicalization (replacing existing words with new ones) overlexicalization (creating an excessive amount of overlapping terms for concepts important to the anti-society), and borrowing from outside the mainstream language, in addition to formal simplification, taboo language, and registerial blurring and cross-over (Lefkowitz & Hedgcock, 2016). It thus mainly exploits lexical and phonological resources to create a variety that is opaque to outsiders.

Examples of anti-language-like systems include the argot of thieves and prisoners, characterized by abundance of words denoting policemen and women (Halliday, 1976). Another example is the French language game Verlan (derived from *l'envers* meaning "backwards" through syllable inversion) originating among multicultural banlieue youth in suburban Paris housing projects, creating a space for enacting aspects of identity distinct from dominant French culture (Lefkowitz & Hedgcock, 2016). It centers around metathesis (inverting syllables, as in *branché* becoming *chébran* for "cool"), combined with relexicalization (*arabe* becomes *beur*, indexing a specific Arab-French youth identity), and overlexicalization (including iterated reverlanization as *beur* becomes *rebeu*).

Notably, anti-languages should not be understood as all-or-nothing categories but as an ideal type which particular varieties may approximate to different degrees (Lefkowitz & Hedgcock, 2016). What matters is that linguistic forms are selected and reworked in order to create distance from a socially salient out-group, while simultaneously strengthening the speaker's alignment with a desired in-group. This makes anti-languages a useful analogue for the phenomenon investigated in this thesis: the strategic language adaptation to signal human as opposed to LLM authorhip can be construed as a form of oppositional identity signaling, and the resulting language variety can be described as an anti-language, which we will call "anti-LLM-ese". The out-group in this case is not a social group of people, but LLMs, whose writing style is increasingly salient and stigmatized. The in-group, by contrast, is the community of human writers who seek to differentiate themselves from LLMs and signal their authenticity through their linguistic choices.

One important distinction between the two should be noted, however. Traditional anti-languages are defined with respect to anti-societies seeking secrecy from the mainstream and solidarity among their members through the development of a shared code. In the case of anti-LLM-ese, however, the large size of the in-group and the contexts of language use, requiring transparency to a large audience and adherence to genre norms, often prevent the possibility of collaborative and dynamic code formation.

The Continuous Cycle of Adaptation

When suspected LLM authorship is costly and reliable arbitration is unavailable, language itself emerges as a site of authentication work. An obvious strategy for such human authorship signaling (hereafter human-authenticating strategy) is to exploit the negative space of LLM-ese by avoiding linguistic features associated with LLM-generated text and emphasizing those associated with human writing.

Such strategies seem to be already in use. For example, when academic publications pointed out that words like "delve" were being overused by ChatGPT in early 2024, their frequency of use measurably decreased in subsequent arXiv abstracts (Geng & Trotta, 2025; Leiter et al., 2024), while other ChatGPT-favored words like "significant," highlighted in less influential papers, continued to increase (Geng & Trotta, 2025). This suggests a strategic avoidance of certain words that have become associated with LLM-generated text. Direct attestations of such behavior appear in public discourse with people sharing they have stopped using the infamous "delve" (AskingAboutChatGPT, 2026) or the em dash (Ice_Efficient, 2026; Si1Fei1, 2025) due to their overwhelming use by LLMs. Some even

share lists of "AI tropes" (Chaib, 2026) and "AI telltale signs" (Fink, 2025) to avoid in writing.

A related strategy is the deliberate introduction of imperfection as a marker of humanness. Recent reports describe writers, including students, as intentionally adding errors or "dumbing down" prose (Al-Sibai, 2025; B. Hu, 2026; Showemimo, 2025). This is striking given that university guidance now warns instructors to be suspicious of work that is "atypically correct in grammar, usage, and editing" ("AI Writing Detection", no date). Lexical innovation is also possible. Lau et al. (2026) reinterpret the "brain rot" slang in this way, away from the mainstream interpretations of cognitive decline in Generation Alpha and instead as a "social creativity strategy" (p. 7). It preserves in-group human identity against an AI out-group by introducing contextual, fast-changing, and absurd expressions, with which AI struggles to keep up. However, this strategy may not be applicable in contexts that require a more formal, codified, or traditional style and that target a wider audience, as opposed to informal interactions among peers.

The last two forms of differentiation, however, are constrained by the need to remain intelligible and to comply with genre conventions and audience expectations. In academic writing, for example, writers cannot introduce too many errors, too much slang, or other highly marked features without undermining communicative success. Another constraint is the solitary situation in which the differentiation takes place, whereby an individual writer tries to prove their authenticity to a sceptical audience without the possibility of coordinating around a shared code in advance.

Even if these strategies prove successful in the short term, they would need to be continuously adapted to keep up with the evolving capabilities of LLMs. New LLMs trained on the most recent data available, would learn to mimic these strategically crafted human texts. Writers would thus be forced to continuously innovate to maintain distinctiveness, producing a cycle of adaptation and counter-adaptation between human writing practices and LLM output. This invites the question of whether humans can sustain this cycle, how they might go about it, and what it implies for the trajectory of the language. The investigation that follows is motivated by these questions.

2.2.3 Iterated Learning as a Framework to Simulate the Evolution of Anti-LLM-ese

Investigating any of the factors and mechanisms reviewed above requires evidence of language as it changes. The most established approach to study language change draws on the historical and corpus-linguistic record. It consists in comparing how a speech community used language at different points in time, either through real-time data, i.e., the same community's productions sampled at separated moments in history, or by inferring ongoing change from synchronic differences across age cohorts within a single community at a single point in time on the assumption that older and younger speakers represent successive stages of an unfolding shift (Millar & Trask, 2023). This is the approach taken, in different ways, by the studies of lexical frequency dynamics, register formation, and word overuse already discussed above (Bochkarev et al., 2014; Degaetano-Ortlieb & Teich, 2018; Kobak et al., 2025; Kousha & Thelwall, 2026), and it has the advantage of tracking change as it actually occurred, in naturalistic productions, without the artificiality that besets more controlled methods. Its main limitation, however, is that it can only ever observe the outcome of change, the shifting frequencies and forms left behind, while the cognitive and interactional processes that produced them must be inferred rather than directly observed.

A second approach formalizes hypothesized mechanisms of change, selection pressures, transmission biases, patterns of contact, as explicit computational or mathematical models, then tests whether

their simulated output reproduces patterns found in real corpora, for example through phylogenetic comparative methods imported from biology to model rates of lexical and grammatical change across language families (Greenhill et al., 2017). This approach can probe the consequences of a mechanism more precisely than the historical record alone, but it does so at the cost of assuming, rather than directly observing, that the mechanism in question is the one actually operative in human speakers.

A third approach instead recreates the conditions of language transmission directly, under controlled experimental conditions, allowing the relevant processes to be observed as they happen rather than reconstructed after the fact. The most established paradigm of this kind is Iterated Learning, in which a sequence, or chain, of simulated agents (Kirby, 2001) or human participants (Beckner et al., 2017; Cornish et al., 2017; Kirby et al., 2008) each learn an artificial communication system from a subset of the output produced by the previous link in the chain, then produce their own output for the next learner in turn. This repeated cycle of partial exposure, learning, and re-production constitutes a transmission bottleneck that filters out idiosyncratic, hard-to-learn variants and progressively favors forms that are more learnable and, often, more compositionally structured, providing a striking laboratory demonstration of how structure can emerge from the cumulative effect of cultural transmission alone, without being designed or directly selected for (Smith et al., 2003).

This thesis adopts the third approach as it allows the process of interest to be observed directly under a controlled setting rather than reconstructed from its eventual traces or simulated under assumptions that themselves require justification. This is particularly relevant given that the phenomenon under investigation, iterative oppositional adaptation against LLMs, is, if real, still in a too early stage to have left a clear signature in the historical record. At the same time, the present study departs from the classic Iterated Learning paradigm in that it is not concerned with how linguistic structure emerges from communicative and transmissioin constraints during language learning. It repurposes the paradigm's iterative transmission of a linguistic signal within a chain-and-generation experimental structure, in which one participant's output becomes the next participant's input, to instead simulate strategic, goal-directed linguistic adaptation under adversarial pressure from a simulated evolving counterpart. This design will serve as the basis for the experimental investigation of the research questions that drive this study, which are detailed below.

2.3 Research Questions and Hypotheses

The considerations presented in this chapter motivate the three research questions we pose:

RQ1 To what extent can individual writers iteratively adapt their writing to signal human authorship in an adversarial setting where LLMs in turn absorb those adaptations, individuals cannot coordinate to create new conventions, and writers must remain intelligible within existing language norms?

RQ2 To what extent do the explanations for adaptations made to signal human authorship change across iterations of this process?

RQ3 To what extent do linguistic properties of the resulting written text, in particular lexical diversity, syntactic complexity, and orthographic, grammatical and punctuation correctness, change across iterations of this process?

With respect to RQ1, the body of literature on socially motivated linguistic adaptation reviewed above suggest that such adaptation could plausibly be sustained, at least for a limited number of

iterations. At the same time, two constraints distinguish this setting from those precedents and may narrow the space of available adaptations. First, writers cannot coordinate with one another to negotiate new shared conventions, as speakers of an anti-language would, and second, any adaptation must remain intelligible and consistent with the conventions of the genre being written in.

With respect to RQ2, a plausible strategy, and one already attested informally, is to target the negative space of LLM-ese, that is, avoiding features associated with LLM output while emphasizing those associated with human writing. Critically, what counts as such a feature is expected to shift over time as LLMs are simulated to absorb prior human-authenticating strategies, so that a feature only remains useful as a differentiation target for as long as writers continue to perceive it as machine-associated. This in turn depends on writers' awareness of the relevant cues, an idea closely related to the access and awareness conditions on successful identity performance discussed above. Whether such awareness tracks the current, evolving state of simulated LLM output or instead reflects writers' prior, possibly outdated beliefs about what LLM-ese looks like has direct consequences for the trajectory predicted. Tracking the current state would favor a more oscillatory, reactive pattern of change, whereas reliance on prior, fixed beliefs would favor a more steadily directional one. A related expectation follows from the tendency for more salient and consciously noticed features to be more available for deliberate manipulation than more abstract ones (Greenhill et al., 2017). Surface-level, easily nameable lexical features (individual overused words) are expected to be targeted earlier in the process than subtler, less salient ones, such as syntactic structure. Notably, as schematically illustrated in Figure 2.1, the features writers target need not coincide with the features that actually, empirically distinguish LLM output from human writing; to the extent that folk perceptions of LLM-ese diverge from its objective linguistic correlates (Jakesch et al., 2023), the strategies captured by participants' explanations may track perceived rather than real cues.

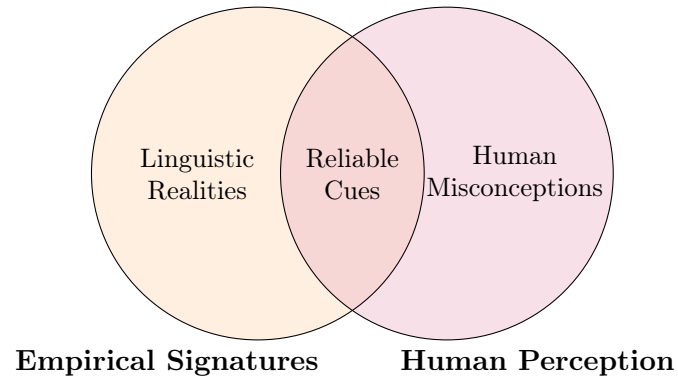


Figure 2.1: Aspects of the differences between LLM-generated and human-writen text. The reliable cues could be exploited in developing strategies for human authorship authentication.

With respect to RQ3, two competing considerations motivate different predictions. On the one hand, if humanness is taken to reside disproportionately in the rare, idiosyncratic "tails" of language (Vanmassenhove, 2025), differentiation strategies might be expected to draw on precisely these tails, favoring out-of-distribution vocabulary and greater variability in syntactic complexity, which would manifest as an increase in lexical diversity and extreme (very high, very low, or very variable) sentence complexity over generations. On the other hand, if writers converge on a similar, limited repertoire of strategies for achieving this effect, the resulting texts might instead become more uniform with one another even as each individual writer intends to differentiate their own text from LLM output, a risk B. Hu (2026)

raises in warning that widespread strategic imitation of "safe," human-coded patterns could flatten writing into a homogeneous style of its own.

Chapter 3

Methods

This chapter introduces the experimental design, the data collection process the data analysis plan. The study was approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam. The study was pre-registered in OSF (https://osf.io/f4pku) and the code and the data are available at https://github.com/raya-mez/machine-speech-human-speech.git.

3.1 Study Design

To simulate the continuous advancement of LLMs in emulating human writing and the resulting pressure on human writers to adapt their strategies to differentiate themselves, we designed an online human-subjects experiment in which participants iteratively edit a text to make it read more convincingly as human-written. The design is inspired by the design of classic Iterated Learning experiments (e.g., Beckner et al., 2017; Kirby, 2001; Kirby et al., 2008). Participants are organized into *chains*, that is, independent sequences of participants, each consisting of up to five *generations*, or sequential positions. Within each chain, the output produced by one participant serves as the input for the next, that is, each participant edits the text edited by the participant in the previous generation. Each chain is initialized with a distinct seed text (see section 3.3 for details), which constitutes the target text for the participant at Generation 1.

This yields a hierarchical structure in which generations are nested within chains and with each chain-generation position occupied by exactly one unique participant in the analyzed dataset (see section 3.4 for details on how this one-to-one mapping is ensured in practice). Chains constitute the independent higher-level units of replication, while observations within chains are non-independent by design as each participant's output directly determines the next participant's input.

Unlike standard Iterated Learning experiments, the number of chains in the present design is not a fixed number but capped with a possibility for early termination. A chain terminates early if two consecutive participants judge their edits unsuccessful in improving human-likeness; otherwise, it continues until the maximum of five generations is reached. This setup was chosen as the degree of sustainability of the hypothesized iterative adaptation process, that is, whether participants would continue finding effective identity-signaling strategies across generations, was unclear and constituted an outcome of interest in itself, targeted by RQ1. We assumed that failure to improve the human-likeness of the text at two successive generations is indicative of a persistent ceiling rather than a temporary lapse, making further data collection uninformative. The available resource could instead be allocated to collecting more independent chains, which was prioritized over greater transmission depth to better

estimate between-chain variation (see section 3.4 for details on sample size considerations).

3.2 Procedure

The experiment takes the form of an online survey containing three sections: a background questionnaire, the main task consisting of three components, and a post-task questionnaire.

Informed consent and use of generative AI. Before taking part in the experiment, all participants provide informed consent. The instructions section of the informed consent form includes a statement about the prohibition of using generative AI tools during the experiment. The importance of this requirement is emphasized by asking participants to explicitly consent to it in a separate question immediately after the informed consent form, before they can proceed to the rest of the experiment.

Background questionnaire. Once consent is provided, participants are asked to confirm they match the demographics required for participation (see section 3.4). If participants do not meet all of the participation requirements, they are automatically redirected to the end of the survey. If participants do meet the requirements, they are then asked about their experience and beliefs regarding LLMs, specifically how often they use LLM (on a five-point scale ranging from "I have never used an AI chatbot / LLM" to "Every day or almost every day"), their view on the future impact of LLMs on society (on a five-point scale ranging from "very detrimental" to "very beneficial," with an additional "no opinion" option), and whether their own writing style has changed after becoming acquainted with LLMs (on a three-point scale from "not at all" through "somewhat" to "a lot"). A first attention check is administered among these questions.

The main task. The main task begins by placing participants in a cover scenario motivating the upcoming text-editing task shown in Figure 3.1.

> Imagine you find yourself in the following situation.
>
> As part of a group project, your team has to submit a 200-word news article and your teammate, Brandon, is in charge of writing it. To do so, he used the newest LLM, designed to have a very human-like writing style.
>
> However, this puts your group at risk. Your professor actively tracks new developments in LLMs and has already cautioned against their use in coursework. You are worried he will suspect the article was produced with the help of an LLM, which would penalize the entire group.
>
> **Your task** is therefore to edit the text so that it reads convincingly as human-written and is unlikely to be identified as LLM-generated.
>
> To get a better sense of the writing style of this new LLM, you first generate two sample articles using it. They are shown below.

Figure 3.1: Cover scenario for the main task.

The cover scenario intentionally designates a human evaluator, namely the professor, rather than an

automated detection system as the imagined judge of text authenticity because humans and AI detection systems do not rely on the same cues to identify LLM-generated text, and are accordingly not equally sensitive to the same differentiation strategies. Automated detectors, for instance, are susceptible to mechanical interventions such as replacing characters with Unicode homoglyphs (Creo & Pudasaini, 2025), which are imperceptible to a human reader. Human judges, in contrast, rely mainly on factual, linguistic and stylistic heuristics (e.g., Georgiou, 2026; Russell et al., 2025), some of which may not accurately reflect the objective differences between human-written and LLM-generated text (Jakesch et al., 2023). The editing task is therefore inherently a theory-of-mind exercise: participants must model what a human reader is likely to perceive as LLM-generated and adjust their text accordingly, a process that operates on folk beliefs about LLM writing rather than on ground truth. This focus is motivated by two considerations. First, if perfect automated detection of LLM-generated text is theoretically intractable (Májovský et al., 2024), convincing human evaluators will remain the relevant challenge in practice. Second, sounding like an LLM is becoming socially stigmatized independently of actual AI use (Al-Sibai, 2025; Geng & Trotta, 2025; B. Hu, 2026): writers may face suspicion on the basis of human heuristics alone, regardless of what any automated system would conclude, and whether the text is actually LLM-generated or not.

Following the cover scenario, two example texts are presented, followed by the target text to be edited (the interface is illustrated in Figure 3.2). After acquainting themselves with these texts, participants complete each of the three components of the main task in sequence: editing, explanation, and assessment of editing success. In the editing component, the target text is displayed again, but sentence by sentence. Each sentence is accompanied by an editable text box prefilled with that sentence, where participants make their edits. All sentences, formatted in this way, are visible throughout the entirely of this phase. The interface is illustrated in Figure 3.3

Once editing is completed, participants proceed to the explanation phase. They are shown all sentences again but with their revisions visually highlighted: additions were shown with a green background, whereas deletions were displayed with red highlighting and strikethrough text (Figure 3.4 shows an example). These visual cues were intended to make the changes more salient, reducing the cognitive load of tracing them back and thus allowing participants to focus on reporting their reasoning. For each sentence they edited, participants are asked to provide a free-text explanation of the changes they made and cannot proceed until they have done so for all edited sentences. To reduce the likelihood of uninformative placeholder responses being submitted, the survey blocks progression until each explanation is at least eight characters long, fewer than half of its characters are digits, and it contains at least two whitespace-separated words, at least two distinct words, and at least one vowel. These explanations were used to inform the actual edits being made and used as proxies for the strategies underlying those edits, which are the focus of RQ2. While free-text explanations are more open to variability and noise than forced-choice strategy questions, they were chosen to avoid biasing participants' responses toward a predefined set of strategies and to allow for the emergence of innovative ones across generations, which was a central aspect of the hypothesized phenomenon of iterative adaptation. The sentence is taken as the unit of analysis, sidestepping the need to define what constitutes a single edit at the character, word, phrase or other level.

Finally, in the last main task component, participants indicate whether they believe their edited version reads more convincingly as human-written than the original text they received ("Yes" or "No"). This questions serves as the early termination criterion mentioned in section 3.1.

EXAMPLE 1

Egypt's parliament approves Red Sea islands transfer to Saudi Arabia

Egypt's parliament has approved a long-debated agreement transferring sovereignty over two Red Sea islands, Tiran and Sanafir, to Saudi Arabia, marking a significant step in a controversial maritime deal. Lawmakers voted in favor of the accord during a plenary session, despite vocal opposition from some MPs and public protests in recent years.

The government has argued that the islands historically belong to Saudi Arabia and that the agreement corrects administrative arrangements dating back to the 1950s. Officials said the transfer would strengthen bilateral ties and unlock economic cooperation, including planned development projects in the Red Sea.

Critics, however, contend that the islands are Egyptian territory and have challenged the move in court. Previous judicial rulings had produced conflicting decisions, fueling a heated national debate over sovereignty and transparency. Security officials emphasized that navigation through the strategic Strait of Tiran will remain open in accordance with international agreements.

Saudi authorities welcomed the vote, describing it as a reaffirmation of historic rights and regional partnership. The agreement now awaits final ratification procedures, after which administrative control is expected to be formally transferred.

Analysts say the decision could reshape regional alliances and test domestic political stability in the months ahead. Across the region.

Figure 3.2: Example texts shown to participants at the beginning of the editing task.

Brandon's text is shown again sentence by sentence, each followed by an editable text box prefilled with that sentence. Make any changes you think will help the text read convincingly as human-written and unlikely to be identified as LLM-generated. You may rephrase, add, or remove content as you see fit. If you choose to merge sentences, enter the combined version in the first relevant text box and leave the following one(s) empty. You can also split sentences by writing multiple sentences within a single text box. Edit as many text boxes as you feel are needed.

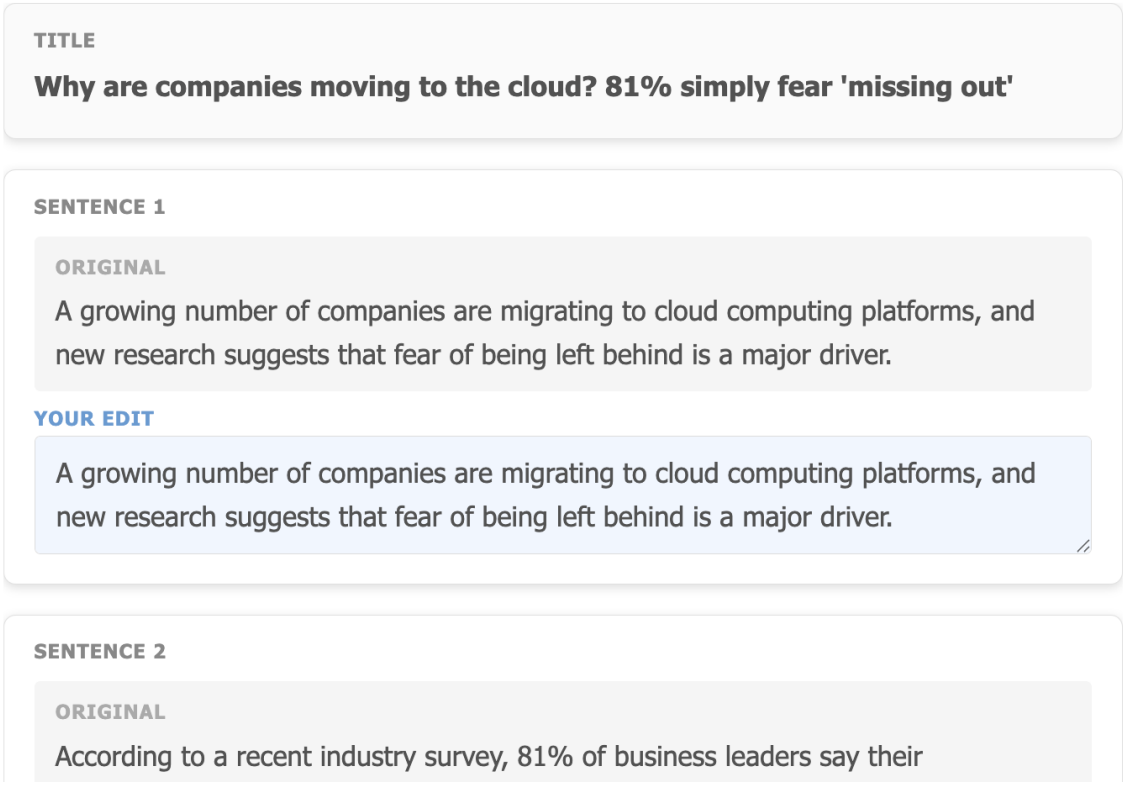


Figure 3.3: Editing task interface.

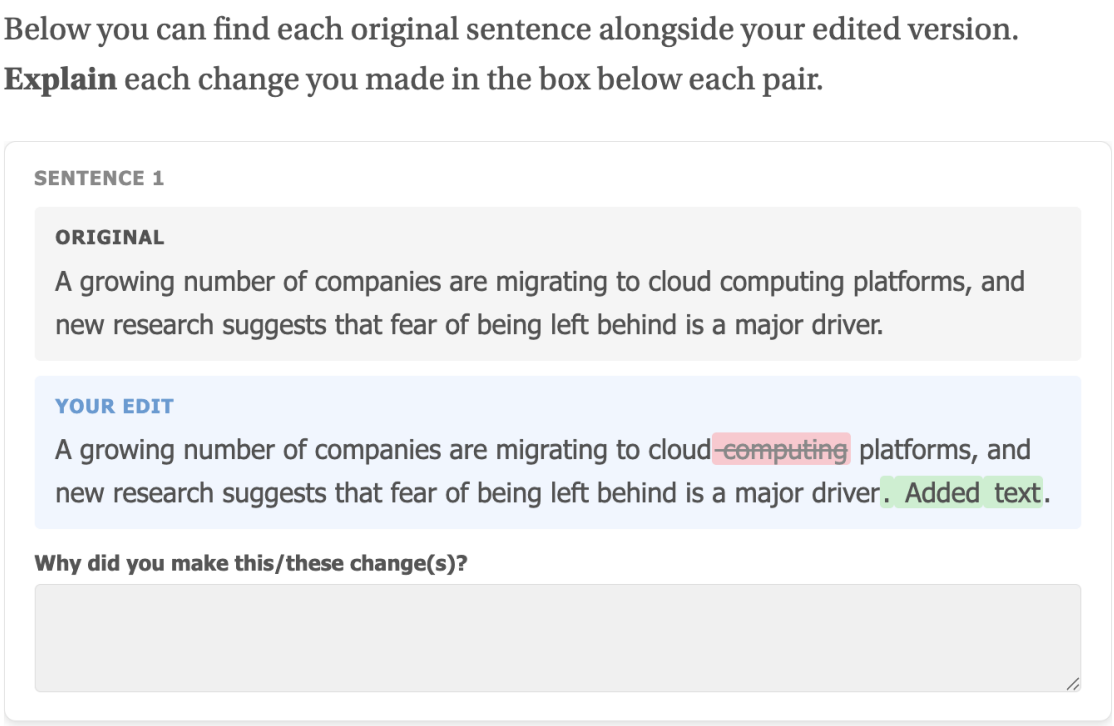


Figure 3.4: Explanation task interface.

Post-task questionnaire. A second attention check is administered after the main task. After that, participants indicate whether they used any generative AI tools during the experiment and whether they expect participation in the experiment to affect their own writing style (if "Yes, if so explain" or "No"). A space for further remarks is available at the end of the questionnaire. The experiment concludes with a debriefing page providing more information about the study and its goals and revealing that the text participants were given to edit might not have been generated by an LLM, contrary to what they were led to believe with contact details provided again.

One survey was created for each generation containing a main task block for each chain. Which block was displayed to a participant was determined through a URL parameter. Thus, a separate link was created for each survey version and the Prolific's Taskflow feature was used to ensure that each link received exactly one response.

3.3 Materials

Each chain is initialized with a distinct approximately 200-word news article generated by GPT-5.2 via the OpenAI API (`gpt-5.2-chat-latest`) using the prompt "Write a 200-word news article titled: [title]. Output only the article text, without the title or any other information." A simple prompt was used to minimize any bias on the writing style of the seed texts. Given that ChatGPT is the most widely used LLM interface (Bick et al., 2026), GPT is the model family whose writing style participants are most likely to have encountered and developed tacit sensitivity to (Russell et al., 2025). Since it is an open empirical question whether the hypothesized iterative adaptation process will emerge at all, using the most familiar model maximizes the chances that participants can identify LLM telltales in the seed texts and act on them, providing the most favorable conditions for the dynamic to manifest if it is present. `gpt-5.2-chat-latest` specifically was used because it was the model powering the

free-tier ChatGPT at the time of method selection, ensuring that the seed texts reflect the writing style participants were likely to have been recently exposed to. Additionally, the GPT model family is the most extensively studied in the literature on human versus LLM-generated writing (e.g., Berriche & Larabi-Marie-Sainte, 2024; Fiedler & Döpke, 2025; Herbold et al., 2023; Opara, 2025; Sandler et al., 2025; Yakura et al., 2025), which facilitates comparison of participants' editing strategies and the linguistic changes they produce with prior findings on perceived and objective differences between human and LLM writing.

News articles were chosen as the target genre because they do not require domain knowledge, their writing style is familiar to a general adult population, and they are fact-based, which encourages participants to focus their editing on linguistic and stylistic cues rather than on thematic content or expressed opinions. They can also be short, self-contained, and topically diverse, making them well-suited to an online survey setting where task duration must be kept manageable and participant engagement maintained. Finally, they have been found to be least affected by LLM-induced linguistic homogenization compared to Reddit posts and scientific articles (Sourati et al., 2025), making them a conservative test case: if human-authenticating editing strategies and their linguistic traces are detectable here, they are likely to be at least as pronounced in genres more heavily affected by LLM use.

Article titles were selected by manual inspection from the CC-News dataset (Hamborg et al., no date). Random entries were drawn until 21 suitable titles were found. Titles were retained if they were likely to produce factual, moderately formal prose that did not invite an interview or opinion format, did not concern sensitive or inappropriate subject matter, and were sufficiently diverse in topic to avoid thematic overlap across chains. This selection procedure was intended to reduce between-chain variance in writing style and register while maintaining topic diversity to keep participants engaged with the task. All 21 titles are listed in Appendix A.2.

At each generation, participants are presented with two example texts alongside the target text. These examples are intended to illustrate the writing style of the purported LLM, giving participants a basis for calibrating their edits. At Generation 1, the two example texts are the seed texts of two other chains, drawn randomly without replacement from the remaining 20 chains and fixed as sources for example texts throughout the generations. If one of the chains terminates early, it is replaced by another randomly drawn chain. At subsequent generations, the example texts are generated by GPT-5.2 (`gpt-5.2-chat-latest`) to imitate the editing strategies of the previous participant in the same chain, applied to the input texts at the previous generation of the two assigned chains for this purpose. Specifically, the model is prompted with the previous participant's edits and explanations as illustrative transformations and instructed to infer the underlying strategies and apply analogous edits to a new text as shown below.

Prompt

```
You are an advanced language model.  Humans have edited your output to make it appear
more convincingly human-written and provided explanations for the edits.  You have
learned from this editing.  Now, you will be given a new text to edit in the same way.

OBSERVED EDITS AND EXPLANATIONS
[example_edits]

INSTRUCTIONS
```

```
    • Infer the strategies humans used in the edits above and apply analogous strategies
      to the new text below.
    • You may rephrase, add, or remove content, merge or split sentences.

NEW TEXT
Title:  [title]
[input_text]

OUTPUT
Return ONLY the fully edited text (without the title).  No explanations or annotations.
```

Figure 3.5: Prompt template to used to generate the example texts. The model is provided with the previous participant's edits and explanations as illustrative transformations and instructed to infer the underlying strategies and apply analogous edits to a new text. The placeholders `[example_edits]`, `[title]`, and `[input_text]` are replaced with the actual content for each chain and generation.

3.4 Participants and Data Collection Procedure

3.4.1 Sample Size

We aimed to collect 105 responses, distributed across chains of up to five generations, which corresponds to 21 full chains, or 21 plus an additional chain for every five chains that terminate early (as described in section 3.1). As will be reported in more detail in section 4.1, no chain terminated early, thus we obtained data from 21 chains spanning five generations.

Given the unavailability of previous literature on the phenomenon under investigation, the target sample size and its distribution into chains and generations was informed by previous literature on Iterated Learning. Classic studies typically include 10 generations per chain (Beckner et al., 2017; Cornish et al., 2017; Kirby et al., 2008), with a more variable number of chains, but commonly revolving around 10. For example, four chains were used in Kirby et al. (2008), which was later more thoroughly reanalyzed using 12 chains per condition by Beckner et al. (2017), and eight chains were used in Cornish et al. (2017). In the present study, the distribution of the sample size was shifted in favour of chain multiplicity over generation depth. This was in consideration of the open-ended nature of the task, which admits large variation in participant behaviors and therefore requires a larger number of independent chains to better estimate between-chain variation and improve the iteration-related inferences. Moreover, whether the chains would continue even beyond the first generation was unclear, thus, progression until five generations was estimated as sufficiently indicative of the presence of an iterative adaptation process. Shorter chains of four to five generation are also not unprecedented in the literature on Iterated Learning (e.g., Partington et al., 2025; Thompson et al., 2025).

Simulation-based power analysis

To supplement the above rationale, we conducted a simulation-based power analysis for the planned analysis of explanations (section 3.5.2). The analysis was informed by a preliminary sample of 5 chains spanning two generations ($N = 71$ observations from 10 participants, averaging 7.1 observations per participant, range: 2–9). We fit a linear mixed-effects model of the form `PC2 ~ generation + (1 | chain / participant)` to the pilot data and used the resulting variance components (between-chain

$SD = 4.54$, between-participant $SD = 0.84$, residual $SD = 4.76$, and a generation slope of $\hat{\beta} = 1.32$ PC2 units per generation) as the basis for simulation.

We then generated synthetic datasets matching the planned collection structure of 21 chains across five generations ($N = 105$) and sampled the number of observations per participant from the empirical pilot distribution. For each simulated dataset, we refit the same mixed model and repeated this procedure for 1,000 simulations per effect-size condition. At the full effect size from the preliminary data, estimated power was 100%. Power reached approximately 80% at about 30% of the pilot effect ($\hat{\beta} \approx 0.40$) and exceeded 85% at 50% of the pilot effect ($\hat{\beta} \approx 0.66$), even under a conservative rule that counted singular fits as failures. Across these conditions, the singular-fit rate was approximately 12%.

We also evaluated Type I error under the null hypothesis using 2,000 simulations. The empirical false-positive rate was 5.15% with a 95% confidence interval of [4.24%, 6.23%], which is consistent with the nominal $\alpha = .05$ level. When singular fits were treated as failures, the estimated Type I error rate was 4.2% (95% CI: [3.38%, 5.20%]).

The empirical Type I error rate under the null ($\beta = 0$, 2,000 simulations) was 5.2% (95% CI: [4.2%, 6.2%]), consistent with the nominal α level of .05. Because the primary analysis involves five principal components tested simultaneously, we additionally simulated power under Holm correction across all five outcomes.

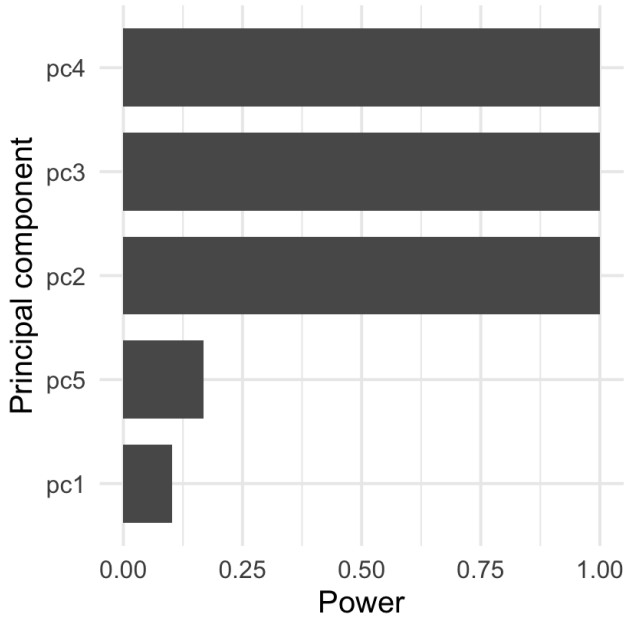


Figure 3.6: Estimated Holm-corrected power from 1,000 simulations for each principal component. Longer bars indicate stronger evidence of detecting the generation effect.

Using 80% of the pilot effect sizes, Holm-corrected power was larger or equal to 99% for PC2, PC3, and PC4 (which carried the largest generation effects), while PC1 and PC5 showed lower power (≈10–17%), reflecting their near-zero generation effects in the pilot data (Figure 3.6). The family-wise probability of rejecting at least one hypothesis was 100% in all configurations tested. Finally, a sensitivity analysis varying the chain-to-generation ratio across a fixed sample size of approximately 105 participants confirmed that concentrating observations into more chains rather than more generations maximizes power for detecting a generation-level trend under the assumed variance structure, supporting the decision to favour chain multiplicity over generation depth.

Because the primary analysis tests five principal components simultaneously, we additionally simulated Holm-corrected power across all five outcomes. At 80% of the pilot effect sizes, Holm-corrected power was at or near 100% for PC2, PC3, and PC4, which showed the largest generation effects in the pilot data. In contrast, PC1 and PC5 had substantially lower power, approximately 10%–17%, reflecting their near-zero generation effects. Across all simulated configurations, the family-wise probability of rejecting at least one hypothesis was 100%.

3.4.2 Participant Recruitment

Participant eligibility and compensation. Participants were recruited through the online participant pool platform Prolific ("Prolific", 2026). They were screened for age (18 or older; mean = 39.3, SD = 14.3, range = [18, 81]) and first language (English). Only native speakers were eligible to participate to reduce confounds related to language background and proficiency. English was chosen as the target language for consistency with the still burgeoning literature on human versus LLM writing, which has focused on English (e.g., Berriche & Larabi-Marie-Sainte, 2024; Fiedler & Döpke, 2025; Herbold et al., 2023; Opara, 2025; Sandler et al., 2025; Yakura et al., 2025).

Participants were compensated at an hourly rate of £8.70, corresponding to £2.90 per session based on an estimated task duration of 20 minutes. This rate was set as close as possible to the maximum permitted by the ethics policy of the Faculty of Humanities at the University of Amsterdam (€2.50 per 15 minutes).

Recruitment procedure and inclusion criteria.

Responses were collected one generation at a time to manage the dependencies inherent to the chain-generation structure, namely the sequential transmission of target texts and the construction of example sets from responses across chains. This also ensured that unusable responses could be replaced before the chain progressed by preventing the compromization of the dependent data in subsequent generations. To ensure that all responses were given by unique participants given this stage-wise recruitment procedure, an additional Prolific screener excluded participants who had already partaken in a previous iteration of the study.

The survey was implemented in Qualtrics and distributed one generation at a time through Prolific's Taskflow feature. While online crowdsourcing platforms, such as Prolific, have been shown to yield data of comparable quality to traditional lab-based methods for linguistic studies (Munro et al., 2010), this advantage has been compromised by the availability of LLMs to participants and of bots completing surveys instead of participants. In response, the platforms have developed automatic systems to detect potential inauthentic responses ("Fraud Detection - Qualtrics", no date; "How Prolific Detects Bots and AI in Online Research", no date), motivating their use for this study. We relied only on the automatic bot detection system of Qualtrics, which assigns a bot likelihood score (Q_Recaptcha) and which uses Google's invisible reCaptcha technology, to each response, and on manual inspection of responses flagged as potentially inauthentic to identify and exclude any responses that were likely produced by bots.

A total of 106 participants were recruited and compensated. One additional participant was collected to replace a unusable response from a participant in Generation 3, who did not fulfill requirement (4) from the following preregistered inclusion criteria:

1. None of the three main components of the main task (editing, explanation, and perceived success) had missing data;

2. The participant passed at least one of the two attention checks;

3. The participant indicated that they did not use any generative AI tools during the experiment;

4. If no edits were made, the participant responded "No" to the question about whether their edits made the text read more convincingly as human-written. Since the question presupposes that the text was at least minimally altered, the opposite answer thus suggests likely inattention or lack of engagement with the task;

5. The response was not flagged as potentially inauthentic by Qualtrics' bot detection system, or if it was flagged, further inspection of the response did not indicate that it was likely produced by a bot. Responses with a Q_Recaptcha score below 0.5 were flagged as potentially inauthentic and excluded if further inspection showed indications of bot-like behavior, such as nonsensical or placeholder responses, or unusually fast response times given the task requirements. The preregistered response time threshold of 12 minutes was relaxed in light of observed response times among participants with a perfect Q_RecaptchaScore score. This led to the retention of one Generation 3 response with a Q_RecaptchaScore of 0.4 and a response time of 9.7 minutes, whose edits and explanations were coherent although not particularly elaborate. This response time was also not implausibly fast given that a participant from the same generation who made edits had a response time of 8.8 minutes.

3.5 Analyses

The data generated by the iterative editing design permit a multitude of complementary analyses at different levels of description. The primary goal of the analysis in this study was to test whether editing behavior and textual properties changed systematically across generations. To that end, we combined a small set of preregistered confirmatory analyses, motivated by prior literature and the research questions, with exploratory follow-up analyses aimed at interpreting and contextualizing the observed patterns more thoroughly.

Notably, because texts are transmitted sequentially within chains, successive generations represent dependent stages in an evolving process rather than independent snapshots. Given this dependence structure and the limited number of observations at each generation compared to the richness of the free-text data, the analyses focused on systematic change across generations rather than on inferring stable strategic or linguistic states at any single point in a chain. All confirmatory models therefore treated generation as the primary predictor of interest and accounted for within-chain dependence using mixed-effects models.

3.5.1 Analysis of Sustainability

Before analyzing changes in editing explanations or text features across generations, we first evaluated how sustainable the iterative editing process was across generations. If most chains were to terminate early, this would suggest that participants quickly exhaust effective differentiation strategies and that the hypothesized iterative adaptation process is not sustainable over time. A lack of early termination would indicate that participants are able to keep finding effective differentiation strategies (based on self-assessment) across multiple generations, providing preliminary evidence for the presence of the hypothesized iterative adaptation process.

A one-sample t-test comparing mean chain length, i.e,. the final generation reached before termination, against the planned maximum of 5 generations was preregistered. However, this test was not theoretically aligned with the nature of chain length as a discrete, bounded outcome. The t-test assumes a continuous, normally distributed variable, which does not match the data-generating process for chain termination. A survival-style or geometric model of termination probability across generations would be more appropriate than a t-test. However, in the actual data, all chains reached Generation 5, so chain length showed no variance (all values equal to 5), making any inferential test undefined, thus none was performed.

Additionally, we conducted qualitative analyses of the cases in which the early termination criterion was partially met, i.e., individual instances in which a participant judges their edits as unsuccessful (without a consecutive second unsuccessful judgment triggering termination). These cases were examined descriptively to understand whether participants made no edits or applied edits but perceived them as ineffective, and to explore any patterns in their responses, including response times and task engagement.

3.5.2 Analyses of Explanations

Confirmatory Analysis of Explanations

To examine whether participants' reported editing rationales changed across generations, we analyzed the free-text explanations provided for each edited sentence. As argued in section 3.2, these explanations were treated as metacognitive proxies for the strategies participants used to signal human authorship. Since the studied process is one of iterative adaptation, novel strategies may emerge across generations based on the edits made by previous generations. They would thus be difficult to anticipate based on prior literature, making it hard to define a coding scheme a priori. One natural candidate would be a taxonomy based on linguistic levels (e.g., lexical, syntactic, or discourse-level changes); however, given the free-text nature of the data, applying such a scheme reliably would require manual annotation by multiple independent experts, which was not feasible given the time and resource constraints of the project. We therefore opted for an data-driven approach, using automatic methods to identify the major dimensions of variation in the explanations.

Concretely, all explanations were first embedded using the `all-mpnet-base-v2` sentence-transformer model (Reimers & Gurevych, 2019). This maps each free-text explanation to a high-dimensional numerical vector (a *semantic embedding*) such that semantically similar explanations are located closer together in the vector space (i.e., the *embedding space*).

To identify the main structure of variation in the explanations, we then applied principal component analysis (PCA) on the obtained embeddings (Jolliffe, 2002). PCA identifies orthogonal linear directions in the embedding space along which the embeddings vary most. Each resulting axis, called a *principal component* (PC), constitutes a new, latent, continuous variable that summarizes one dominant way in which the explanations vary. Projecting the embeddings onto these axes yields a set of *PC scores* for each explanation, indicating where it falls along each PC. We retained the scores for the five PCs accounting for the most variation in the dataset for the subsequent analyses of generation-dependent changes in explanations. This allowed us to limit the number of statistical tests that need to be corrected for multiple comparisons, thereby preserving statistical power, while capturing the most dominant patterns of variation.

PCA was preferred over alternative dimensionality-reduction methods (e.g., t-SNE or UMAP) because these methods are sensitive to hyperparameters and stochastic initialization, making them less suitable

for deriving stable dimensions of variation for confirmatory analyses. PCA, in contrast, is deterministic and yields a fixed set of globally defined components, allowing explanations to be located within a common coordinate system and compared directly using their PC scores.

To assess whether explanations changed across generations along the identified dimensions, we tested the effect of generation on each retained PC using linear mixed-effects models with random intercepts for chain and participant nested within chain:

$$\mathrm{PC}_k \sim \mathrm{generation} + (1 \mid \mathrm{chain/participant}), \quad k = 1, \ldots, 5. \tag{3.1}$$

Participant-level random intercepts account for the non-independence of multiple explanations produced by the same participant (one per sentence edited), while chain-level random intercepts account for dependencies induced by the iterative transmission of texts within chains. Although the explanations themselves are not transmitted across generations, they reflect edits made to texts that are. For example, if one participant introduces and comments on a spelling irregularity, that alteration becomes part of the text inherited by the next participant and may shape the subsequent edits and explanations. Generation was centered at 1 ($\mathrm{generation}_{\mathrm{centered}} = \mathrm{generation} - 1$) such that the intercept corresponds to Generation 1.

A significant effect of generation on a given PC indicates a systematic linear increase or decrease along that dimension over time; a null effect suggests no such trend is observed. For example, if a PC captures the degree to which explanations reflect lexical (higher scores) as opposed to syntactic edits (lower scores), a positive effect of generation would indicate explanations becoming more lexically- and less syntactically-oriented over time, whereas a negative effect would indicate the opposite pattern. Note, however, that this is an illustrative example and the actual interpretation of the PCs is not known a priori, which motivates the first exploratory analysis described below.

Models were fitted using the `lme4` package (Bates et al., 2015) in R, with p-values obtained via the Satterthwaite approximation as implemented in `lmerTest` (Kuznetsova et al., 2017). Model assumptions were evaluated by visual inspection of residual and Q-Q plots. No substantial violations were detected. Singular fits and convergence issues, when present, are reported.

Exploratory analyses of explanations

The confirmatory analyses of PC scores were complemented by three sets of exploratory follow-up analyses.

First, because PCA was applied to dense sentence embeddings whose dimensions are not directly interpretable, the resulting PCs do not come with straightforward semantic labels. We therefore inspected explanations with the highest and lowest scores on each component to infer the dimensions of semantic variation each component appeared to capture.

Second, upon visual inspection of the distribution of PC scores across generations, we observed a potential nonlinear pattern in the scores of some of the PCs. Therefore, we fitted supplementary models including a quadratic term for generation to test for quadratic trends across generations that might have been obscured in the linear models. These models were otherwise identical to the linear models described above:

$$\mathrm{PC}_k \sim \mathrm{generation} + \mathrm{generation}^2 + (1 \mid \mathrm{chain/participant}), \quad k = 1, \ldots, 5. \tag{3.2}$$

Finally, we estimated covariate-adjusted models to examine whether individual-level demographic and attitudinal predictors (age, LLM usage frequency, perceived societal impact of LLMs, and perceived effect of LLMs on one's writing style) accounted for variation in the outcomes. These models also followed the same structure as the linear models described above, with the addition of the covariates as fixed effects:

$$\begin{aligned} \mathrm{PC}_k \sim\ & \text{generation} + \\ & \text{age} + \\ & \text{LLM usage} + \\ & \text{perceived societal impact of LLMs} + \\ & \text{perceived effect of LLMs on own writing style} + (1 \mid \text{chain/participant}) \end{aligned} \tag{3.3}$$

LLM usage frequency and perceived societal impact of LLMs were modelled as ordered categorical variables. Perceived societal impact of LLMs were included as a categorical predictor and coded using treatment contrasts, with "No impact" serving as reference categories.

3.5.3 Analyses of Text Features

Whereas the analysis of editing strategies addresses what participants consciously report doing, the analysis of linguistic text features addresses whether those interventions leave systematic, measurable traces in the texts themselves.

Confirmatory Analyses of Text Features

We selected lexical diversity, syntactic complexity, and correctness as the three main features of interest for confirmatory analysis because they recur in the literature on perceived differences between human and LLM-generated writing (Georgiou, 2026; Russell et al., 2025) and in informal reports of strategic text altering for human authorship authentication already taking place (Al-Sibai, 2025; B. Hu, 2026).

Feature operationalization. Lexical diversity was operationalized as the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis (2010)), computed using the `LexicalRichness` Python package. MTLD tracks the running type-token ratio (TTR) through a text sequentially, counting the number of segments, or *factors*, required before the TTR falls below a threshold standardly fixed at 0.72, and divides the total number of tokens by the mean number of factors across a forward and backward pass. Higher values indicate that a text sustains lexical variety over longer stretches before repetition reduces the TTR. This operationalization was chosen over alternative lexical diversity measures following previous literature (Herbold et al., 2023; McNamara et al., 2010; Mizumoto et al., 2024; Muñoz-Ortiz et al., 2024), which favors it over simple TTR as it is less sensitive to text length.
Syntactic complexity was operationalized as mean dependency distance, computed using the `TextDescriptives` Python package (Hansen et al., 2023) with the `en_core_web_lg` spaCy model. It is computed as the average linear distance in words between syntactically linked tokens across all dependency relations in a text. Higher values indicate that syntactic dependencies span longer distances on average, reflecting greater structural complexity.
Correctness was operationalized as the total count of grammar, spelling, punctuation, and casing alerts flagged by the open-source proofreading software LanguageTool ("LanguageTool", no date) via its

Python wrapper `language_tool_python` (Jack Morris, 2026). To avoid penalizing legitimate variation between American and British English spelling conventions, only alerts flagged by both the `en-US` and `en-GB` models were retained. Alerts in the `STYLE` category were excluded, as these reflect stylistic preferences rather than objective errors. Raw alert counts rather than per-word rates were used as the outcome variable, with text length controlled for via a log word count offset in the statistical model.

Statistical models. Mixed-effects models were fit in R using `lme4`, `lmerTest`, and `glmmTMB`, depending on model type.

For each of lexical diversity and syntactic complexity, a linear mixed-effects model was fitted with the feature value as the outcome, generation as a fixed effect, and a random intercept for chain:

$$\text{feature} \sim \text{generation} + (1 \mid \text{chain}). \tag{3.4}$$

Chain is included as a grouping factor to account for the non-independence of observations within the same chain, as each participant's output directly determines the next participant's input. Generation was centered at 1 ($\text{generation}_{\text{centered}} = \text{generation} - 1$) such that the model intercept corresponds to Generation 1.

For correctness, we fit a generalized linear mixed-effects model with the raw number of error alerts in each text as the outcome and a Poisson distribution following the recommendations of Winter and Wieling (2016) for dealing with count data. To control for text length, we included an offset term of the log word count. This setup effectively models the log rate of errors per word. For comparability with prior work reporting error rates per 100 words (e.g., Mizumoto et al., 2024), descriptive statistics are reported on that scale. As in the other models, we used generation as the fixed effect and included a random intercept for chain:

$$\text{error count} \sim \text{generation} + (1 \mid \text{chain}) + \text{offset}(\log(\text{word count})). \tag{3.5}$$

For the linear mixed-effects models, assumptions were evaluated by visual inspection of residual and Q-Q plots. For the Poisson mixed-effects model, we additionally assessed dispersion using simulation-based diagnostics in `DHARMa`. No substantial violations were detected.

Exploratory analyses of text features

These confirmatory analyses were complemented by three sets of exploratory follow-up analyses. First, we estimated covariate-adjusted models to examine whether individual-level demographic and attitudinal predictors (age, LLM usage frequency, perceived societal impact of LLMs, and perceived effect of LLMs on one's writing style) accounted for variation in the outcomes. Second, for text features whose association with generation remained statistically significant after correction for multiple comparisons, we conducted additional sensitivity analyses to assess the robustness of the observed effects to alternative model specifications. Finally, for these features, we conducted descriptive follow-up analyses to characterize the observed effects in greater detail. Specifically, we examined within-chain differences in the feature relative to the immediately preceding generation and relative to the seed text.

3.5.4 Statistical inference

For all models, statistical significance was evaluated using two-tailed tests as there was no strong a priori prediction about the direction of change across generations. To control the family-wise error rate across the eight planned confirmatory comparisons, p-values were adjusted using Holm's sequential correction (Holm, 1979) and compared against the standard $\alpha = 0.05$ threshold. Holm's method was preferred over the standard Bonferroni correction because it retains greater statistical power while still controlling the family-wise Type I error rate. While the Bonferroni correction applies a fixed threshold of $\frac{\alpha}{m}$ (m being the number of comparisons) to all tests, Holm's procedure is stepwise: p-values are first ordered from smallest to largest, and each is compared to a progressively less stringent threshold of $\frac{\alpha}{m-k+1}$, where k denotes the rank of the p-value. Once a comparison fails to reach significance, all larger p-values are treated as non-significant. This sequential procedure still controls the family-wise error rate because each step is calibrated so that the chance of making even one false rejection across the whole family remains bounded by α, but it is less conservative than Bonferroni because not every test is forced to meet the same hard threshold α/m.

Chapter 4

Results

This chapter presents the results of analyses addressing the three research questions: (1) the sustainability of iterative human-authorship authentication, (2) change in editing explanations across generations, and (3) change in text features across generations, alongside exploratory analyses informing the interpretation of the results. All chains reached the maximum generation, providing support for the sustainability of iterative human-authorship authentication. No evidence emerged for linear change in the content of editing explanations across generations as captured by principal components of explanation embeddings. Regarding text features, we find that lexical diversity decreased across generations, while no significant linear change was observed for mean dependency distance and error rate after correction for multiple comparisons.

4.1 Sustainability of Iterative Human-Authorship Authentication

We assessed the sustainability of iterative human-authorship authentication based on the proportion of chains that reached the last, fifth, generation, as opposed to terminating early due to two consecutive participants judging their edits as unsuccessful in improving the human-authorship signaling of the text. All 21 chains reached Generation 5, indicating that participants continuously found ways to better signal human authorship, even after multiple prior rounds of editing had already done so. These results support the sustainability of iterative human-authorship authentication, with no evidence of a ceiling effect at least for the first five generations.

While no two consecutive participants within any chain both judged their edits unsuccessful in improving the human-authorship signaling of the text to trigger early termination, five individual participants reported such unsuccessful editing: one participant in Generation 1, one participant in Generation 4, and 3 participants in Generation 5. All five of these cases were from distinct chains: 15, 8, 7, 13, and 18, respectively. The first two participants (Generation 1 in chain 15 and Generation 4 in chain 8) and one of the latter three (Generation 5 in chain 13) made no edits to the text, suggesting they possibly considered the input text as sufficiently convincingly human-authored or they did not find any means to edit it. Alternatively, they may not have deeply engaged with the task. While we cannot determine the exact reason for their making no edits, the latter explanation is supported by the fact that these three participants had short response times ($M = 11.3, SD = 4.0$), corresponding to approximately half of the overall mean ($M = 22.2, SD = 12.7$). The response times of these three participants were also shorter than the response times of the other two participants in Generation 5, who reported unsuccessful editing despite applying edits ($M = 28.9, SD = 11.5$). A remark left by the participant in Generation

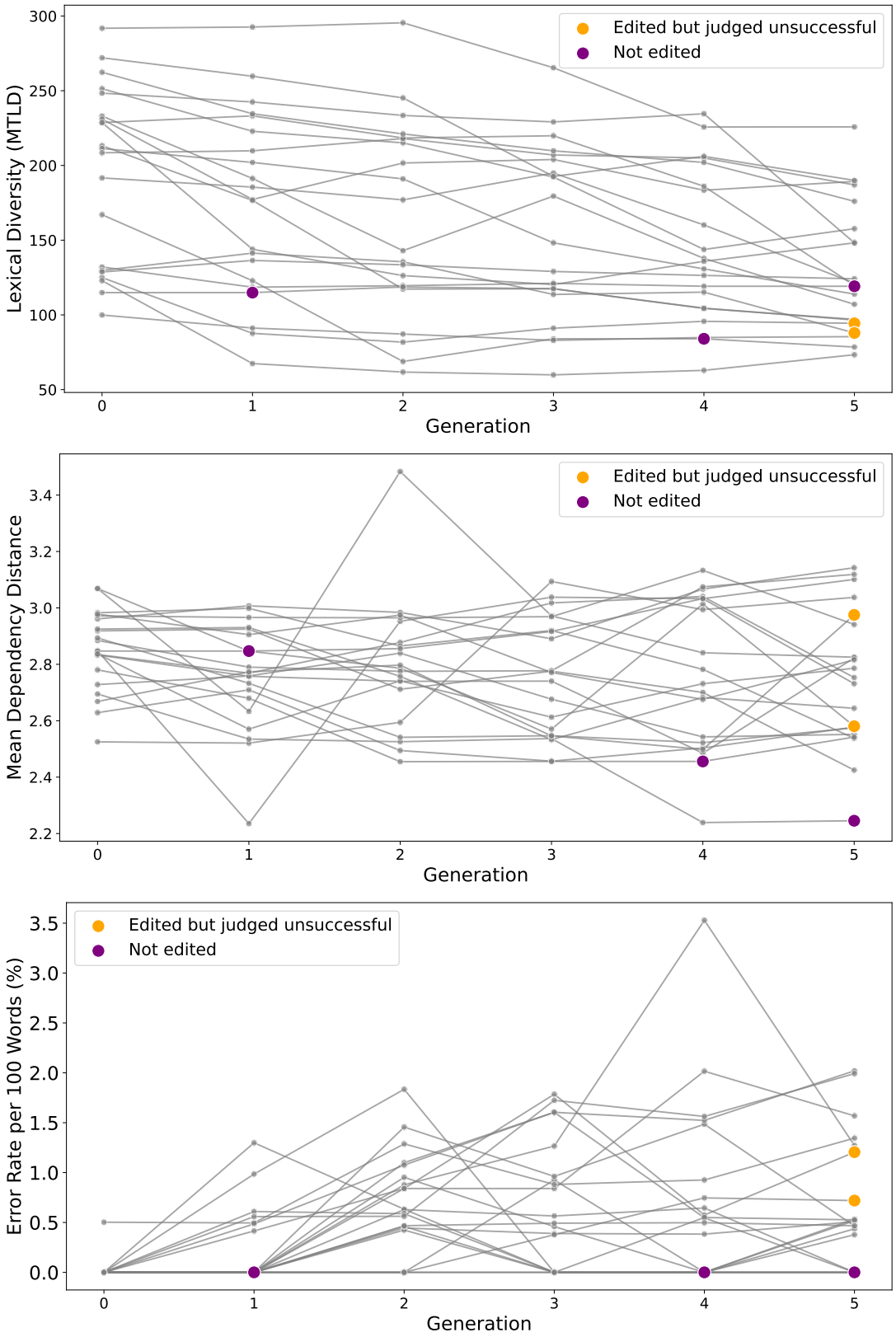


Figure 4.1: Measure of Lexical Diversity (MTLD), mean dependency distance, and error rate across generations for each chain, highlighting cases of self-assessed unsuccessful editing. Gray lines represent individual chains. Orange points indicate texts that were edited but whose authors reported the editing as unsuccessful, and purple points indicate texts that were not edited. Across all three measures, these cases fall within the broader range of observations, providing no clear indication that self-assessments of unsuccessful editing were associated with unusually high or low feature values.

1 stating: “I thought the generated text was formal and suits the context of the article, which suits the grammar of the AI model.” further shows a misunderstanding of the task as aiming to *match* the writing style of the purported LLM rather than to *differentiate* the target text from it. However, this seemed to be an isolated case of such misunderstanding.
Exploratory inspection of the two cases in which participants applied edits but nevertheless judged them unsuccessful did not reveal a clear common profile in the text features of interest. As shown in Figure 4.1, the texts produced by the concerned participants fell within the broader observed range for MTLD, mean dependency distance, and error rate, rather than appearing as clear outliers. Relative to the three other cases in which participants reported no improvement in human-authorship signaling because they made no edits, the unsuccessfully edited cases showed some differences in mean dependency distance and error rate, but these patterns were not consistent enough across measures to suggest a distinctive feature profile. Given the very small number of cases, these observations should be treated as descriptive only.

4.2 Change in Editing Explanations across Generations

To assess whether participants’ reported editing rationales changed across generations, we applied PCA to the semantic embeddings of their free-text explanations and tested for linear effects of generation on the scores of the first five principal components (PCs).
Because the principle components are latent dimensions that capture data-driven axes of variation in a semantic vector space rather than predefined semantic constructs, we first inspect what each component appears to capture and assign it a descriptive proxy label. This will facilitate interpretation of the statistical results by providing a more intuitive, albeit tentative and subjective, way to conceptualize and refer to the components when presenting and discussing the findings.

4.2.1 Explanations with Extreme PC Scores

To build an intuition for the dimensions of variation captured by the PCs, we review the five explanations with the highest and the five with the lowest scores on each component. We label each component using an “x vs y” format that reflects the two poles of the distribution. The first three extremes are shown in Table 4.1; the full set is available in Appendix B.1.

PC1: transformation-focused vs problem-focused rationales High PC1 scores seem to be attributed to explanations that relate lexical substitution or paraphrasing aimed at increasing perceived human-likeness (e.g., “replaced words to make it sound more human” or “changed words and added some to make it sound more human”). Low PC1 scores, by contrast, seem to be associated with terse justifications that refer directly to the content of the sentence being edited (e.g., “The domestic party scenes are the majority.” or “They don’t fuel any[t]hing.”). These explanations omit the editing action itself and instead point to a perceived issue in the original sentence, such as an inaccurate characterization, an unsupported claim, or an inappropriate formulation. For example, the explanation “The domestic party scenes are the majority.” accompanying the replacement of “The spectacle is fueled by advertisers” with “The spectacle is fueled by domestic party scenes” suggests that the participant regarded the original metaphor as inaccurate or misleading. Such explanations are often only fully interpretable in light of the sentence being edited. PC1 therefore appears to distinguish decontextualized,

PC	Highest-scoring explanations	Lowest-scoring explanations
PC1	replaced words to make it sound more human	The domestic party scenes are the majority.
	changed words and added some to make it sound more human	They don't fuel anyrhing.
	simplified/changed words to seem more human written	Everyone who watches, so not nearly.
PC2	it sounds more natural and humanlike	Wrong punctuation - combined the sentences.
	It seems more human	Merged the sentences and removed a part which I felt was not needed.
	It sounded more human.	The sentence was too long and needed some punctuation to make it easier to read
PC3	Seems more natural.	Too perfect punctuation is a first giveaway that text wasn't written by a human. LLMs like to add too many examples. LLMs tend to always use American English spelling.
	This sounds more natural.	removing extra words a human may not have used
	Sounds more natural	because AI's don't make simple typos like this
PC4	To change the structure. Starting with the statement captures attention, who said it follows because it's not always relevant information.	Better word
	I felt this was a fair sentence to keep as it gives a nice point in the timeline without making every sentence too wordy.	Spelling error
	I thought this information was important in the opening statement - answering the who, what, where, when, why questions.	the spelling wasn't right
PC5	To make the sentence flow better.	I kept the factual content but made the tone more casual and reduced formal wording like "major charitable."
	To make the sentence flow more naturally, in accordance with how a real human talks.	I made a couple of changes that could make the text more natural sounding without changing the meaning very much.
	To make it flow a little better	This one was short so it was hard to change it too much, but it sounds a bit more natural to have "unknown".

Table 4.1: Explanations whose embeddings had the highest and lowest scores on PCs 1-5. Background colors indicate explanations from the same participant among the top five highest or lowest scoring explanations for at least one of the five PCs. Thus, the explanations with unique colors in this table indicate participants who had an explanation among the top or bottom five, that was not in the top or bottom three shown in this table. The full set of the five highest and five lowest scoring explanations for each PC is available in Appendix B.1.

transformation-focused rationales from source-text critiques that motivate why a revision was deemed necessary.

The color coding in Table 4.1 and Appendix B.1 shows that the two highest-scoring explanations on PC1 originate from a single participant (Generation 2 of Chain 2), the second two from another participant (Generation 5 of Chain 11), and all three lowest-scoring ones are contributed by a third participant (Generation 1 of Chain 5). This suggests that part of the variance captured by the component reflects individual differences in how participants frame their editing actions, and that a given participant tends to apply a similar framing across the several edits made within their own turn. This anticipates the results of the statistical analysis below, illustrating how a substantial share of the variation in editing rationales is better understood as a stable feature of who is writing than as a function of where in the iterative process they are writing. This is worth bearing in mind when interpreting the absence of significant generation effects below.

PC2: global human-likeness judgments vs structural editing operations High PC2 scores seem to be associated with brief, outcome-oriented evaluations of the edited sentence, typically expressed as holistic judgments of human-likeness or naturalness (e.g., "it sounds more natural and humanlike" or "It seems more human"). These explanations do not specify particular linguistic operations, but instead summarize the perceived overall effect of the edit at a global level, often in terms of how "human" or "robotic" the sentence appears. Low PC2 scores, in contrast, are associated with explicit descriptions of sentence-level editing operations and restructuring decisions. These include punctuation corrections, sentence merging, shortening, and content reduction (e.g., "Wrong punctuation - combined the sentences" or "Merged the sentences and removed a part which I felt was not needed"). In addition to describing what was changed, these explanations often justify why structural intervention was necessary, such as improving clarity or consolidating information. PC2 therefore appears to distinguish global evaluative judgments of human-likeness from concrete descriptions of structural editing actions applied at the sentence level. In contrast to PC1, which concerns whether explanations are framed in terms of transformations versus problem diagnoses, PC2 seems to captures the level of abstraction at which participants reason about their edits: from holistic impression to explicit syntactic manipulation.

PC3: intuitive naturalness judgments vs explicit LLM-heuristic reasoning High PC3 scores seem to encode brief, intuitive evaluations of fluency and naturalness (e.g., "Seems more natural" or "This sounds more natural"). These explanations focus on the immediate perceived quality of the edited sentence, without reference to specific linguistic features or explicit reasoning about why the text appears natural. Low PC3 scores, by contrast, seem to be associated with explicit heuristic reasoning about properties of LLM-generated text. These explanations invoke general beliefs about how AI systems typically produce text, such as the absence of typographical variation, overuse of examples, or consistent spelling conventions (e.g., "Too perfect punctuation is a first giveaway that text wasn't written by a human. LLMs like to add too many examples..." or "because AI's don't make simple typos like this"). PC3 therefore appears to separate intuitive judgments of naturalness from explicit metalinguistic heuristics about the expected characteristics of LLM-generated text.

PC4: discourse-level reorganisation vs surface-level correction Explanations with high PC4 scores justify edits in terms of discourse organization, rhetorical structure, or communicative intent, describing why information should be ordered or presented in a particular way (e.g., "To change the

structure. Starting with the statement captures attention..." or "I thought this information was important in the opening statement - answering the who, what, where, when, why questions."). In several cases, participants also refer to genre expectations or perceived norms of LLM-generated text, using these as motivation for restructuring decisions. Low-scoring explanation on PC4, on the other hand, consist of minimal references to surface-level edits such as spelling correction or word substitution (e.g., "Better word" or "Spelling error"). PC4 therefore appears to distinguish explanations grounded in discourse-level organisation and communicative framing from those focused on isolated surface-level corrections.

PC5: underspecified fluency motivations vs elaborated stylistic transformations Finally, high PC5 scores seem to be assigned to brief and highly underspecified explanations that justify edits in terms of general improvements to flow or readability (e.g., "To make the sentence flow better." or "To make it flow a little better"). These explanations typically do not specify what linguistic changes were made, instead relying on a global notion of improved fluency or coherence. In some cases, they also reflect minimal local fixes framed in very general terms (e.g., "to finish the quotation"). Low PC5 scores, in contrast, appear to be associated with detailed descriptions of specific stylistic and lexical transformations applied to the text with the concerned explanations often enumerating multiple changes within a single sentence (e.g., "I kept the factual content but made the tone more casual and reduced formal wording like 'major charitable' " or "Removed the hyphen in 'co-founded' and swapped 'says' for 'has said'... Also changed 'radical overhaul' to 'big overhaul' because radical felt too formal..."). PC5 therefore seems to distinguish minimal, outcome-level fluency justifications from detailed, multi-operation stylistic editing narratives that specify how the perceived improvement was achieved.

Similarly to PC1, this components also reflects some individual-level variance, with the two highest-scoring explanations originating from a single participant (Generation 3 of Chain 18) and two low-scoring explanations originating from another participant (Generation 4 of Chain 3). This again suggests some consistency in how participants frame their editing rationales across generations, in this case in terms of the level of detail and specificity in their explanations.

These descriptive labels should be read as shorthand for the dominant pattern visible at the extremes of each component, not as claims that any PC captures only one aspect of explanation variation. In practice, several components are consistent with multiple features of the explanations, including length, specificity, and the linguistic level targeted. The contrast labels simply foreground the main axis of variation apparent in the highest- and lowest-scoring examples.

4.2.2 Variance Explained by the Five PCs

Taken together, the five PCs explained 19.2% of the total variance in the explanation embeddings (PC1: 5.4%, PC2: 4.8%, PC3: 3.4%, PC4: 2.9%, PC5: 2.7%). This suggests that the explanations were rather heterogeneous, with substantial variation remaining outside the first five components and spread across many dimensions. The scree plot in Figure 4.2 illustrates this, showing the gradual decline in the percentage of variance explained by each component. This is not surprising given the open-ended nature of the task, which can elicit a variety of editing rationales, as well as the free-text format of the data, which allows for idiosyncratic ways of expressing these rationales. The subsequent analysis therefore should be regarded in this context: concerning only the most dominant axes of variation

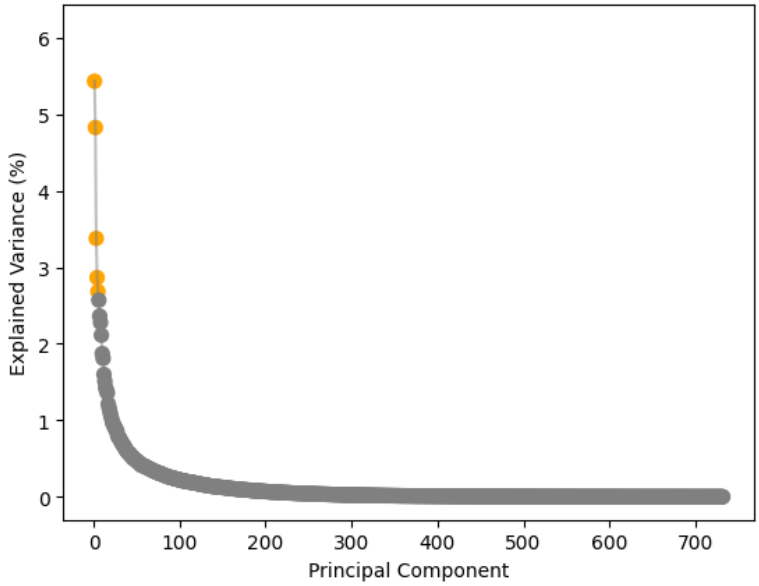


Figure 4.2: Scree plot showing the percentage of variance explained by each principal component (PC) of the explanation embeddings. The first five PCs, which were included in the subsequent analyses, are highlighted in purple. These five components together explain 19.2% of the total variance (PC1: 5.4%, PC2: 4.8%, PC3: 3.4%, PC4: 2.9%, PC5: 2.7%), indicating that while they capture some dominant dimensions of variation in the explanations, a large portion of the variance remains distributed across many other components. Note that the vertical axis reaches only 6% for better visibility, as the first component explains only 5.4% of the variance.

described above, rather than attempting to capture the full complexity of the explanation space.

4.2.3 Confirmatory Analyses: PC Scores Across Generations

To test whether participants' reported editing rationales changed across generations, we modeled the scores on the first five PCs using linear mixed-effects models. The models included a fixed effect of generation (baseline-centered such that the intercept represents the outcome for Generation 1) and random intercepts for participant nested within chain. The results are summarized in Table 4.2 and Figure 4.3 illustrates the corresponding fitted trajectories across generations alongside the observed mean PC scores for each chain at each generation. Inspection of residual and Q-Q plots did not reveal substantial deviations from normality or severe heteroscedasticity, thus no transformations were deemed necessary.

Across all five components, no evidence of a linear effect of generation was found. The estimated effects were small ($|\beta| < 0.012$ across all models) and none reached significance (all raw $p > .15$; all Holm-corrected $p > .92$; see Table 4.2). This indicates that the way participants explained their edits across generations was not found to systematically shift along any of the primary dimensions of variation in the explanation embedding space—they did not, for example, become more focused on transformation-focused rationales as opposed to problem-focused rationales (PC1), or shift from global human-likeness judgments toward structural editing operations (PC2), etc., across generations.

Notably, the random effects structure in the models (Table 4.3) further reveals that variance attributable to participants (nested within chains) consistently dominated that attributable to chains across all components. Participant-level SDs ranged from 0.073 (PC5) to 0.108 (PC1), while the chain-level SD was only half that for PC1 (0.051) and near zero for PC2–PC5. Thus, once participant-level variation

PC	b	SE	t	df	p	p_{Holm}	Sig.
PC1	0.008	0.009	0.84	82.1	.403	1.000	ns
PC2	−0.011	0.009	−1.25	98.4	.211	1.000	ns
PC3	−0.005	0.007	−0.72	76.0	.472	1.000	ns
PC4	0.010	0.007	1.43	79.3	.154	0.923	ns
PC5	−0.004	0.007	−0.60	101.5	.551	1.000	ns

Table 4.2: Fixed effects of generation on the first five principal components (PCs) of explanation embeddings, estimated with linear mixed-effects models. The table reports the estimated coefficient (b), its standard error (SE), the corresponding t statistic, degrees of freedom (df), the nominal p value, the Holm-corrected p value (computed across all eight hypothesis tests in the study), and the significance indicator (with "ns" denoting a non-significant effect at $\alpha = 0.05$ for the Holm-corrected values).
Generation was coded so that the intercept corresponds to Generation 1, and the coefficient b represents the estimated linear change in PC score at each successive generation.
Across all five components, the estimated generation effects were very small ($|b| < 0.012$) and not statistically significant before or after Holm correction, indicating no evidence that participants' editing rationales followed a linear trend across generations along any of the major dimensions of variation in the explanation embedding space, i.e., transformation-focused vs problem-focused rationales (PC1), global human-likeness judgments vs structural editing operations (PC2), intuitive naturalness judgments vs explicit LLM-heuristic reasoning (PC3), discourse-level reorganisation vs surface-level correction (PC4), or underspecified fluency motivations vs elaborated stylistic transformations (PC5).

Component	Group	Variance	SD
	Participant (within chain)	0.012	0.108
PC1	Chain	0.003	0.051
	Residual	0.026	0.162
	Participant (within chain)	0.011	0.104
PC2	Chain	0.000	0.005
	Residual	0.026	0.161
	Participant (within chain)	0.006	0.079
PC3	Chain	0.000	0.012
	Residual	0.020	0.140
	Participant (within chain)	0.006	0.076
PC4	Chain	0.000	0.009
	Residual	0.016	0.126
	Participant (within chain)	0.005	0.073
PC5	Chain	0.000	0.000
	Residual	0.015	0.122

Table 4.3: Random effects variance components and standard deviations (SD) from linear mixed-effects models predicting principal component scores of explanation embeddings, by component (PC1–PC5). Variance is partitioned into three sources: participant (within chain), chain, and residual. Across all components, participant-level variance is substantially larger than chain-level variance, indicating that individual differences between participants drive variation in explanation style more than differences between revision chains. For PC2 and PC5, chain-level variance is near zero, yielding singular fits and suggesting little residual between-chain variation once participant-level variation is accounted for.

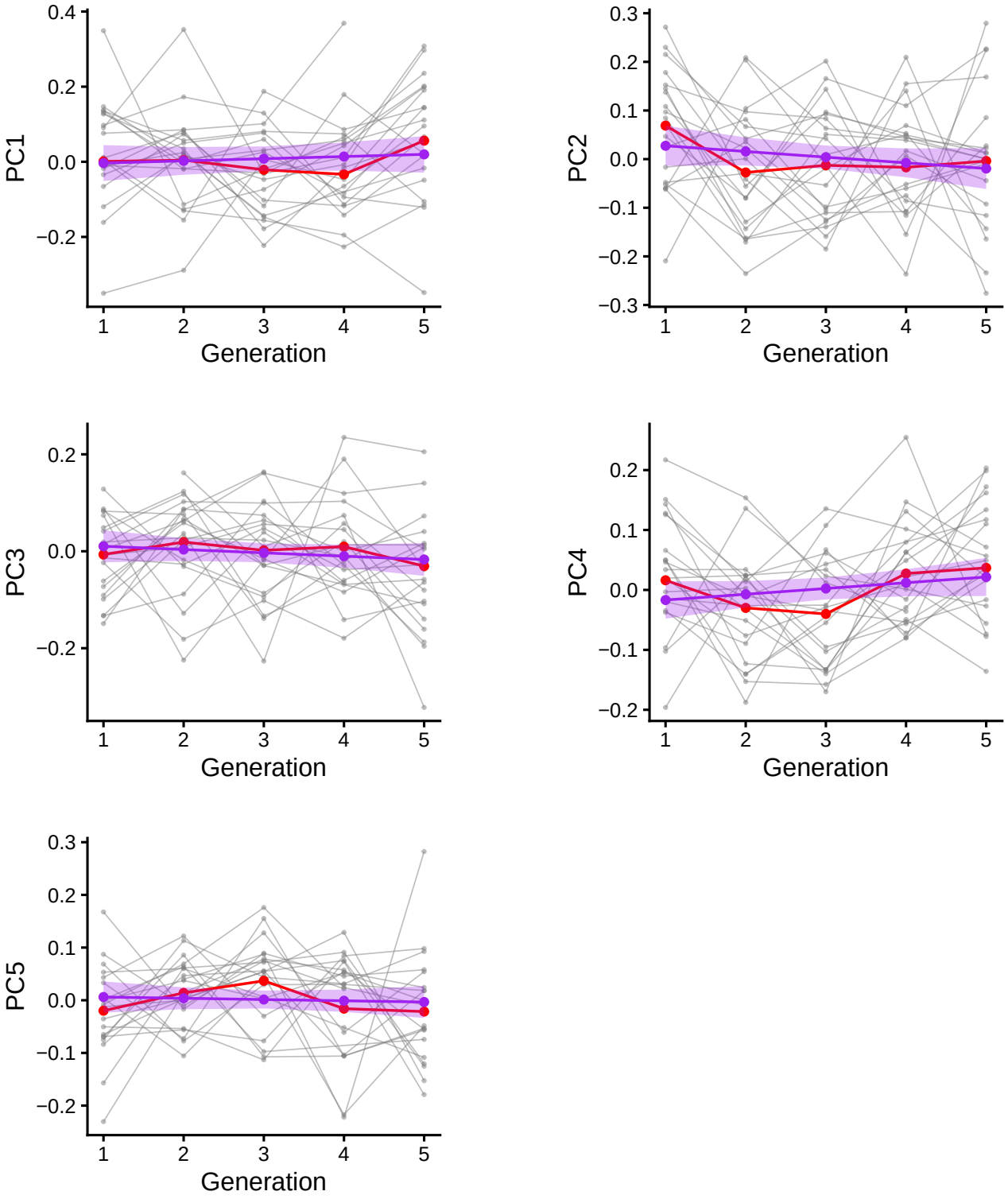


Figure 4.3: Each subplot shows observed and fitted PC scores across generations. The gray points correspond to the mean PC scores for each chain at each generation (i.e., per participant), with lines connecting the same chain across generations. In red are shown the mean scores across chains at each generation. The purple lines show the fixed-effect trajectories of the PC scores across generations estimated from linear mixed-effects models. The shaded bands show 95% confidence intervals around the fixed effect estimates.
The near-flat fitted trajectories indicate no systematic linear increase or decrease across generations along any of the primary dimensions of variation in the explanation embedding space.

was accounted for, there little between-chain variation remained, leading to a boundary (singular) fit for PC5. In other words, this indicates that individual differences between participants drive variation in the content of editing explanations considerably more than differences associated with the particular sequence of revisions applied to the text within each chain.

Additionally, residual SDs exceed both participant-level and chain-level SDs across all components (0.015–0.026, Table 4.3), indicating that a substantial portion of variance in PC scores remains unexplained by the model's fixed and random effects. This is consistent with the relatively low cumulative variance explained by the first five PCs (19.2%), suggesting that much of the variation in explanation embeddings lies in dimensions not captured by these components.

4.2.4 Exploratory Analysis of Non-Linear Effect of Generation

Although the linear mixed-effects models did not reveal a significant effect of generation on any of the five PCs, visual inspection of the trajectories of the observed PC scores (the gray and red lines in Figure 4.3) suggested some components might vary non-linearly across generations. For example, the PC4 scores (discourse-level reorganisation vs surface-level correction) appear to follow a U-shaped trajectory, decreasing from Generation 1 to Generation 3, then increasing until Generation 5. Conversely, the PC5 scores (underspecified fluency motivations vs elaborated stylistic transformations) seem to show the opposite trajectory, increasing from Generation 1 to Generation 3 and decreasing thereafter. This pattern can be obscured by a linear model because changes in opposite directions across the range of generations may cancel each other out in the estimated slope. Therefore, to test whether accounting for this curvature improved model fit, we fit exploratory follow-up models including a quadratic term for generation (baseline-centered such that the intercept represents the scores for Generation 1), with all other specifications matching the linear models (linear term for baseline-centered generation and random intercepts for participant nested within chain).

Table 4.4 reports the quadratic model fit results, and Figure 4.4 visualizes the corresponding estimated trajectories for each principal component. The quadratic term was nominally significant for PC1, i.e., the transformation-focused vs problem-focused rationales component ($b = 0.016$, $SE = 0.007$, $t = 2.201$, $p = .031$), PC4, i.e., the discourse-level reorganisation vs surface-level correction component ($b = 0.015$, $SE = 0.005$, $t = 2.807$, $p = .006$), and PC5, i.e., the underspecified fluency motivations vs elaborated stylistic transformations component ($b = -0.012$, $SE = 0.005$, $t = -2.254$, $p = .026$), suggesting that the trajectories of these components across generations may indeed be non-linear. The positive quadratic term for the former two indicates a U-shaped trajectory, while the negative term for the latter indicates an inverted U-shaped trajectory. Intuitively, this means that over the first generations, explanations shifted increasingly toward problem-focused rationales, surface-level corrections, and underspecified fluency motivations, as opposed to transformation-focused rationales, discourse-level reorganisation, and elaborated stylistic transformations, then this trend reversed in the final generations.

The linear term was nominally significant for PC2, i.e., the global human-likeness judgments vs structural editing operations component ($b = -0.064$, $SE = 0.030$, $t = -2.163$, $p = .033$), PC4, i.e., the discourse-level reorganisation vs surface-level correction component ($b = -0.049$, $SE = 0.022$, $t = -2.247$, $p = .027$), and PC5, i.e., the underspecified fluency motivations vs elaborated stylistic transformations component ($b = 0.044$, $SE = 0.021$, $t = 2.046$, $p = .043$), indicating a net directional change across generations in these components. Notably, these linear effects do not contradict the results from the strictly linear models, which found no significant linear effect of generation on these components.

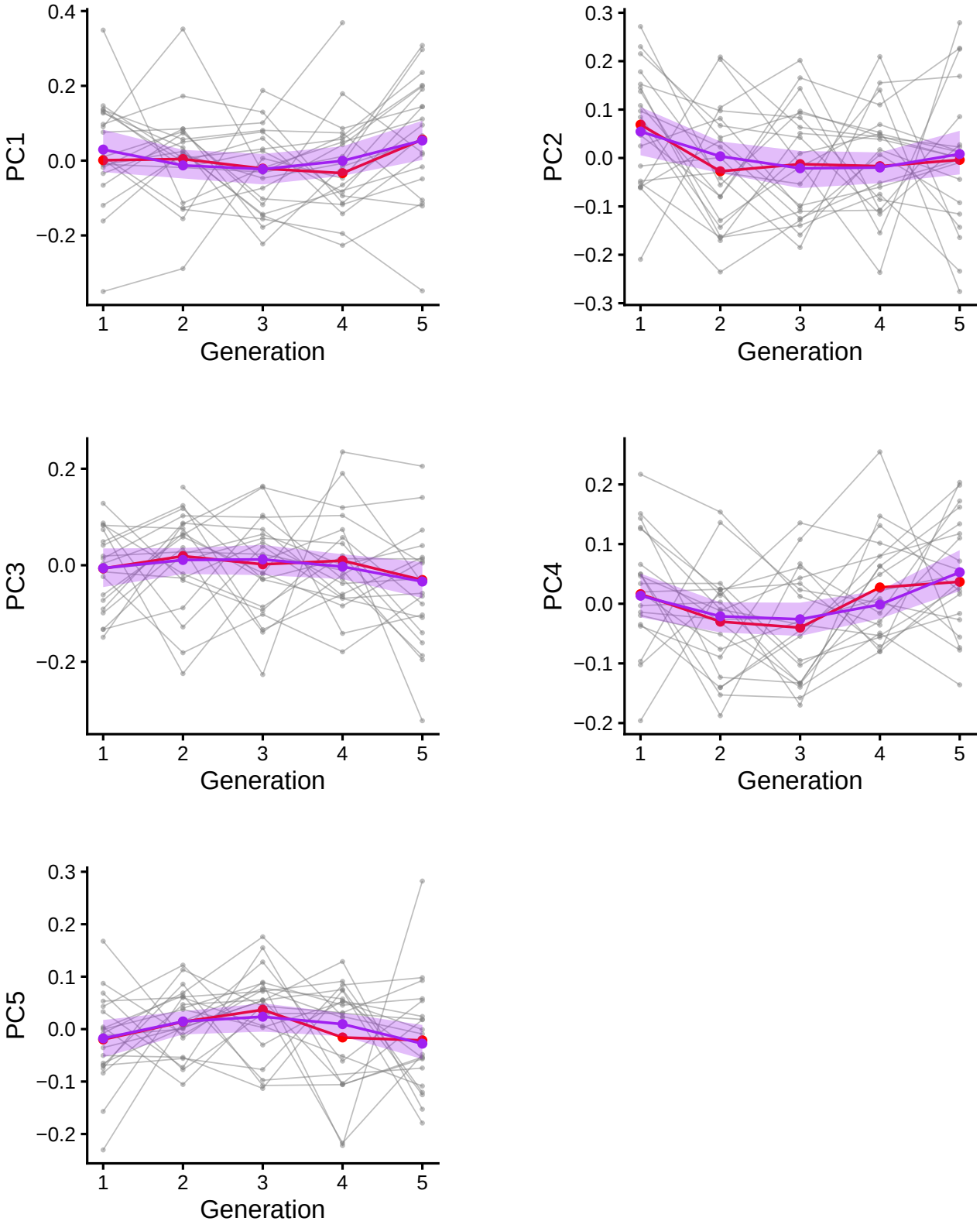


Figure 4.4: Purple lines show the fixed-effect trajectories of PC scores across generations estimated from linear mixed-effects models with a quadratic term for generation. The shaded bands show 95% confidence intervals around the fixed effect estimates. Grey points correspond to the mean PC scores for each chain at each generation (i.e., participant-level means), with lines connecting the same chain across generations. The near-flat trajectories indicate no systematic increase or decrease across generations along any of the primary dimensions of variation in the explanation embedding space.

PC	Term	b	SE	t	df	p	Sig.
PC1	generation_centered	−0.058	0.031	−1.914	80.5	.059	ns
	generation_centered^2	0.016	0.007	2.201	81.0	.031	*
PC2	generation_centered	−0.064	0.030	−2.163	81.5	.033	*
	generation_centered^2	0.013	0.007	1.846	81.9	.069	ns
PC3	generation_centered	0.025	0.024	1.041	74.1	.301	ns
	generation_centered^2	−0.008	0.006	−1.385	74.7	.170	ns
PC4	generation_centered	−0.049	0.022	−2.247	78.5	.027	*
	generation_centered^2	0.015	0.005	2.807	78.9	.006	**
PC5	generation_centered	0.044	0.021	2.046	101.7	.043	*
	generation_centered^2	−0.012	0.005	−2.254	102.1	.026	*

Table 4.4: Results for the quadratic mixed-effects models predicting principal component scores of explanation embeddings, including both the linear and quadratic terms for generation. Generation was baseline-centered such that the intercept represents the expected value for Generation 1. The table reports the estimated coefficients (b), standard errors (SE), t-statistics, degrees of freedom, and p-values for both the linear and quadratic terms. The "Sig." column indicates the nominal significance level of each effect, as this was a follow-up exploratory analysis. The values are "ns" for non-significant, "*" for $p < .05$, and "**" for $p < .01$.
The linear term was significant for PC2, PC4, and PC5, but not for PC1 or PC3. The quadratic term was significant for PC1, PC4, and PC5. It was positive for PC1 and PC4, indicating to a U-shaped trajectory over generations, and negative for PC5, indicating to an inverted U-shaped trajectory.

For PC4 (discourse-level reorganization vs surface-level correction) and PC5 (underspecified fluency motivations vs elaborated stylistic transformations), the unmodeled curvature in the linear models reduced the statistical power to detect the directional trend. This is because when data follows a curved trajectory but is fitted with a straight line, the model attempts to estimate a single slope across the entire range, and the curvature adds unexplained variance that inflates the standard error of the slope estimate. PC2 (global human-likeness judgments vs structural editing operations), on the other hand, showed a significant negative linear effect without a significant quadratic effect. This suggests a weak monotonic trend that became detectable once the quadratic model reduced residual variance by accounting for minor curvature, even though that curvature was itself not statistically significant. Explanations thus became increasingly focused on structural editing operations compared to global human-likeness judgments across generations, although by a small margin.

Conversely, PC1 (transformation-focused vs problem-focused rationales) exhibited a significant quadratic effect but not a significant linear effect. This indicates a symmetric U-shape whereby decreases in the early generations are balanced by increases in later generations, producing no net directional change. Explanations emphasized transformation-focused rationales more than problem-focused rationales in the first generations, then shifted toward problem-focused rationales in the middle generations, and finally returned to emphasizing transformation-focused rationales in the later generations.

For PC4 and PC5 (discourse-level reorganization vs surface-level correction and underspecified fluency motivations vs elaborated stylistic transformations), both terms were significant, indicating curved trajectories that still show an overall directional trend (PC4 decreasing overall, PC5 increasing overall). This means that explanations initially shifted toward surface-level corrections and underspecified fluency motivations and away from discourse-level reorganization and elaborated stylistic transformations, then reversed toward the latter in the later generations. However, the latter trends were still stronger than the former, resulting in a net increase in the focus on discourse-level reorganization and elaborated stylistic transformations with an associated decrease in brief explanations of surface-level corrections and underspecified fluency motivations across generations.

Across all components with significant quadratic effects, the small effect sizes ($|b| < 0.02$), also apparent in the mild curvatures of the fitted trajectories in Figure 4.4, indicate that the non-linear patterns represent modest tendency shifts rather than drastic reversals in the content of the explanations across generations.

The random effects structure for the quadratic models (Table 4.5) is consistent with the linear models, showing that variance attributable to participants (nested within chains) is consistently larger than variance attributable to chains across all components. Participant-level standard deviations (SDs) ranged from 0.005 (PC5) to 0.11 (PC1), while chain-level SD was only 0.003 for PC1 and near zero for PC2–PC5, with PC5 again reaching a boundary (singular) fit. Thus, individual differences between participants accounted for considerably more variation in explanations than differences between editing chains, as was the case in the linear models. Residual SDs were also larger than both participant-level and chain-level SDs across all components, ranging from 0.015 (PC5) to 0.026 (PC1). This suggests that a substantial portion of the variance in PC scores remains unexplained by the model's fixed and random effects, which is consistent with the relatively low percentage of variance explained by the first five PCs (19.2%).

To assess whether the quadratic models provided a better fit than the linear models, we conducted a likelihood-ratio test and computed information criteria (AIC and BIC). The test indicated significantly

Component	Group	Variance	SD
	Participant (within chain)	0.011	0.104
PC1	Chain	0.003	0.055
	Residual	0.026	0.162
	Participant (within chain)	0.0102	0.101
PC2	Chain	0.000	0.013
	Residual	0.026	0.161
	Participant (within chain)	0.006	0.079
PC3	Chain	0.000	0.012
	Residual	0.020	0.140
	Participant (within chain)	0.005	0.072
PC4	Chain	0.000	0.014
	Residual	0.016	0.126
	Participant (within chain)	0.005	0.071
PC5	Chain	0.000	0.000
	Residual	0.015	0.122

Table 4.5: Random effects variance components and standard deviations (SD) from quadratic mixed-effects models predicting principal component scores of explanation embeddings, by component (PC1–PC5). Variance is partitioned into three sources: Participant (within chain), Chain, and Residual. Across all components, participant-level variance substantially exceeds chain-level variance, indicating that individual differences between participants drive variation in explanations far more than differences between revision chains. For PC2, PC4, and PC5, chain-level variance is near zero, suggesting little residual between-chain variation once participant-level variation is accounted for. Residual SDs exceed both participant-level and chain-level SDs across all components, indicating that a substantial portion of variance remains unexplained by the model's fixed and random effects.

improved fit for PC1, i.e., the transformation-focused vs problem-focused rationales component ($\chi^2(\text{df} = 1) = 4.75$, $p = .029$), PC4, i.e., the discourse-level reorganisation vs surface-level correction component ($\chi^2(\text{df} = 1) = 7.77$, $p = .005$), and PC5, i.e., the underspecified fluency motivations vs elaborated stylistic transformations component ($\chi^2(\text{df} = 1) = 5.15$, $p = .023$). However, both AIC and BIC favored the linear model for all five PCs (see Table 4.6), indicating that while the quadratic term captures meaningful curvature for these three components, the improvement in fit is modest relative to the penalty for added complexity and the linear model would be preferred under parsimony principles.

PC	ΔAIC	ΔBIC	LR χ^2	df	p
PC1	5.27	9.86	4.75	1	.029
PC2	6.70	11.30	3.40	1	.065
PC3	8.56	13.16	1.95	1	.163
PC4	2.98	7.58	7.77	1	.005
PC5	5.66	10.26	5.15	1	.023

Table 4.6: Comparison of linear and quadratic mixed-effects models predicting principal component scores from generation. The table reports the difference in AIC and BIC between the two models (ΔAIC = AIC_quadratic − AIC_linear and ΔBIC = BIC_quadratic − BIC_linear), the likelihood-ratio test statistic (χ^2), degrees of freedom, and the corresponding p-value, which was not corrected for multiple comparisons as this was an exploratory analysis. Positive ΔAIC and ΔBIC values indicate that the linear model had the lower information criterion, that is, a better trade-off between fit and complexity. The results suggest that although the quadratic model provided some improvement in fit for a few components, the quadratic term did not improve model fit enough to offset the added complexity, so the linear model was preferred by both AIC and BIC for all five PCs.

Additionally, these results were obtained from exploratory follow-up analyses that were not pre-registered and were not corrected for multiple comparisons. Therefore, they should be interpreted with caution and regarded as tentative evidence for non-linear patterns in the evolution of editing explanations across generations.

In sum, the analyses provide some indications that participants' reported editing rationales changed following a U-shaped pattern across generations along certain dimensions. Specifically, participants initially shifted toward greater focus on problem-focused rationales relative to transformation-focused rationales (PC1), and toward surface-level corrections and underspecified fluency motivations relative to discourse-level reorganization and elaborated stylistic transformations (PC4 and PC5), then reversed toward the latter in later generations, with a net increase in the focus on discourse-level reorganization and elaborated stylistic transformations compared to surface-level corrections and underspecified fluency motivations. However, despite the likelihood-ratio test favoring quadratic models for these three components, AIC and BIC favored the linear models, indicating that the improvement in fit was modest relative to the added complexity.

4.2.5 Exploratory Analysis of Demographic and Attitudinal Predictors

To assess whether individual differences could account for variation in explanation scores, we estimated exploratory covariate-adjusted versions of the linear PC models. The adjusted models included age, LLM usage frequency, perceived societal impact of LLMs, and perceived effect of LLMs on writing style as fixed-effect predictors, alongside the fixed effect of generation and the same random-intercept structure as in the main models. We used the linear models rather than the quadratic models for these analyses because the former were pre-registered, while the latter were exploratory and did not provide

a strongly fit to justify their added complexity.

Across all five PCs, inclusion of these covariates did not alter the conclusions of the confirmatory analyses: the effect of generation remained small and non-significant in every model (Appendix Table B.2). Thus, individual differences did not account for the absence of generation effects.

Most covariate effects were also non-significant, with two isolated associations emerging, but no systematic pattern across predictors. The first concerned PC1 (editing operations vs. source properties) and perceived influence of LLMs on one's own writing style. A marginally significant linear trend ($b = -0.12$, $SE = 0.06$, $t = -1.92$, $p = .058$) indicated that participants who reported that LLMs impacted their writing style "a lot" tended to have slightly lower PC1 scores than those who reported no influence at all, that is, their explanations put a greater emphasis on source properties relative to editing operations. The quadratic term, however, was significant ($b = -0.085$, $SE = 0.041$, $t = -2.09$, $p = .039$), indicating that the relationship between perceived LLM influence and PC1 scores was not strictly monotonic. Specifically, participants reporting that LLMs affected their writing style "somewhat" did not fall proportionately between the two endpoint groups, suggesting a more complex relationship than a simple linear trend.

Secondly, participants who reported having no opinion regarding the societal impact of LLMs had significantly higher PC2 scores (overall human-likeness judgments vs. structural editing operations) than participants who believed that LLMs will have no impact on society ($b = 0.244$, $SE = 0.108$, $t = 2.25$, $p = .027$). However, all other societal-impact contrasts were non-significant, suggesting the observed effect was an isolated result exploratory rather than a general attitude effect.

4.2.6 Summary of Results for Editing Rationales

To summarize, the semantic embeddings of participants' free-text explanations of their editing rationales revealed five relatively interpretable dimensions. These seemed to capture contrasts such as transformation-focused versus problem-focused rationales (PC1), global human-likeness judgments versus structural editing operations (PC2), intuitive naturalness judgments versus explicit LLM-heuristic reasoning (PC3), discourse-level reorganisation versus surface-level correction (PC4), and underspecified fluency motivations versus elaborated stylistic transformations (PC5). Together, these five components accounted for just under 20% of the variance in the explanation embedding space, indicating largely idiosyncratic reasoning reports, consistent with their free-form nature.

The preregistered linear models found no evidence that PC scores varied across generations. However, follow-up models suggested modest quadratic patterns for PC1, PC4, and PC5, indicating that participants' reported editing rationales may have first shifted toward one pole of the PC continua (toward problem-focused rationales, surface-level corrections, and underspecified fluency motivations), then reversed toward transformation-focused rationales, discourse-level reorganisation, and elaborated stylistic transformations. Covariate-adjusted models suggested only two isolated exploratory associations, but no meaningful effect of demographic and attitudinal variables on the PCs. Taken together, these findings do not provide reliable evidence that editing rationales shifted systematically across generations along the dimensions captured by the PCs, despite tentative signs of non-linear change in a few components. The next and final section of the results will examine whether the interventions themselves, nevertheless, led to systematic changes in the linguistic features of the texts across generations.

4.3 Change in Text Features across Generations

The following analyses examine whether the reported interventions analyzed in the previous section are visible in the texts themselves as systematic, measurable linguistic changes. In particular, they will focus on lexical diversity as measured by the Measure of Textual Lexical Diversity (MTLD), syntactic complexity as measured by mean dependency distance, and error rate as measured by the number of errors relative to word count. The confirmatory analyses of generation effects on these three text features are presented first, followed by exploratory analyses.

4.3.1 Confirmatory Analyses

Among the three confirmatory text features, lexical diversity showed the clearest evidence of change across generations. The linear mixed-effects model found that MTLD decreased significantly ($b = -10.40$, $SE = 1.33$, $t(83) = -7.82$, $p < .001$) and this effect remained significant after Holm correction ($p_{\text{Holm}} < .001$), indicating that texts included a narrower vocabulary (i.e., fewer unique words) as generations progressed. As shown in Figure 4.5, this downward trend was largely consistent across chains, although not at a uniform rate. Chains starting with higher MTLD (around 250–290) tended to drop more steeply, while those starting lower (around 115–130) tended to decline more gradually or plateau. This pattern is reflected in the large between-chain variance (Table 4.8) captured by the random intercepts (SD = 52.63), which was substantially larger than the residual variance (SD = 19.27). In practical terms, this means that most of the variation in MTLD scores across the dataset reflects which chain, and therefore which starting point, a text belongs to, rather than how a text happens to deviate from its own chain's trajectory at a given generation. We return to, and formally test, the apparent relationship between starting MTLD and rate of decline in the random-slopes sensitivity analysis reported at the end of this section.

By contrast, mean dependency distance showed no evidence of systematic change across generations ($b = -0.005$, $SE = 0.012$, $t(83) = -0.39$, $p = .699$). The fitted line in the corresponding plot in Figure 4.5 is flat and the individual chain trajectories either fluctuate around it or remain relatively stable across generations, without showing a consistent pattern. This is reflected in the random-effects structure (Table 4.8) with only modest between-chain variation (SD = 0.140) compared to the residual one (SD = 0.177), indicating that most variability occurs within chains rather than as stable differences between them. Intuitively, this is close to the opposite pattern from the one just described for MTLD. Chains are not reliably distinguishable from one another in their typical level of syntactic complexity, and a text with a relatively long mean dependency distance at one generation is about as likely to be followed, within the same chain, by a shorter one as by another long one. Mean dependency distance therefore behaves more like noise fluctuating around a single, shared average across the whole dataset than like a feature with either a systematic generational trend or stable, chain-specific baselines.

The error rate showed a positive trend across generations but did not reach statistical significance after Holm correction ($b = 0.167$, $SE = 0.064$, $z = 2.62$, $p = .009$, $p_{\text{Holm}} = .061$). Notably, Figure 4.5 shows that the mean error rate across chains increased through Generations 1 and 2 up to approximately 0.7 errors per 100 words and then remained relatively stable across subsequent generations. Thus, later generations overall contained more errors relative to word count than the seed texts, but the increase was confined to the early generations rather than continuing over time. The plot also reveals non-negligible between-chain variability with most chains remaining near zero throughout, whereas

others reaching 1.5–2 errors per 100 words, and one chain spiking to over 3 errors at generation 4. This pattern is consistent with the chain-level random intercept variance (SD = 0.649), corresponding to nearly two-fold differences in expected error counts on the response scale.

Feature	b	SE	t/z	df	p	p_{Holm}	Sig.
MTLD	−10.40	1.33	−7.82	83.0	$< .001$	$< .001$	***
Mean dependency distance	−0.005	0.012	−0.39	83.0	.699	1.000	ns
Error count (offset for log word count)	0.167	0.064	2.62	—	.009	.061	ns

Table 4.7: Results for the fixed effect of generation on the three text features of interest, with Holm-corrected p-values computed across all eight planned comparisons. Significance levels: $p < .05$ (*), $p < .01$ (**), $p < .001$ (***), ns = not significant. The test for error count was performed with a generalized linear mixed-effects model with a log link function and an offset for log word count, thus the test statistic is a z-value rather than a t-value and no degrees of freedom are reported.

Component	Group	Variance	SD
MTLD	Chain (intercept)	2770.200	52.630
	Residual	371.400	19.270
Dependency distance	Chain (intercept)	0.019	0.140
	Residual	0.031	0.177
Error count (Poisson log scale)	Chain (intercept)	0.422	0.649
	Residual	—	—

Table 4.8: Random effects variance components and standard deviations (SD) from mixed-effects models predicting MTLD, mean dependency distance, and error counts across chains. All models include a random intercept for chain. Variance is partitioned into between-chain variation (chain intercepts) and residual variation. MTLD shows large between-chain heterogeneity relative to residual variance, dependency distance shows minimal between-chain clustering, and the Poisson error model shows moderate chain-level variability on the log scale. For the Poisson model, residual variance is not separately estimated because observation-level variance is determined by the mean–variance relationship of the Poisson distribution.

Inspection of residual and Q-Q plots did not reveal substantial deviations from normality or severe heteroscedasticity, and no transformations were performed.
To sum up, these three confirmatory analyses show an unequal effect of generation on different text features. Lexical diversity declined robustly across generations, while syntactic complexity and error rate showed little or weaker evidence of systematic change. Before drawing substantive conclusions from this asymmetry, however, it is worth asking whether this may be attributable to individual differences among participants. The next subsection addresses this by re-estimating the three models with demographic and attitudinal covariates included.

4.3.2 Exploratory Analyses with Individual-Level Predictors

To assess whether individual differences among participants could account for variation in the feature scores of the produced texts, we additionally estimated an exploratory mixed-effects model including demographic and attitudinal covariates (age, LLM usage frequency, perceived societal impact of LLMs,

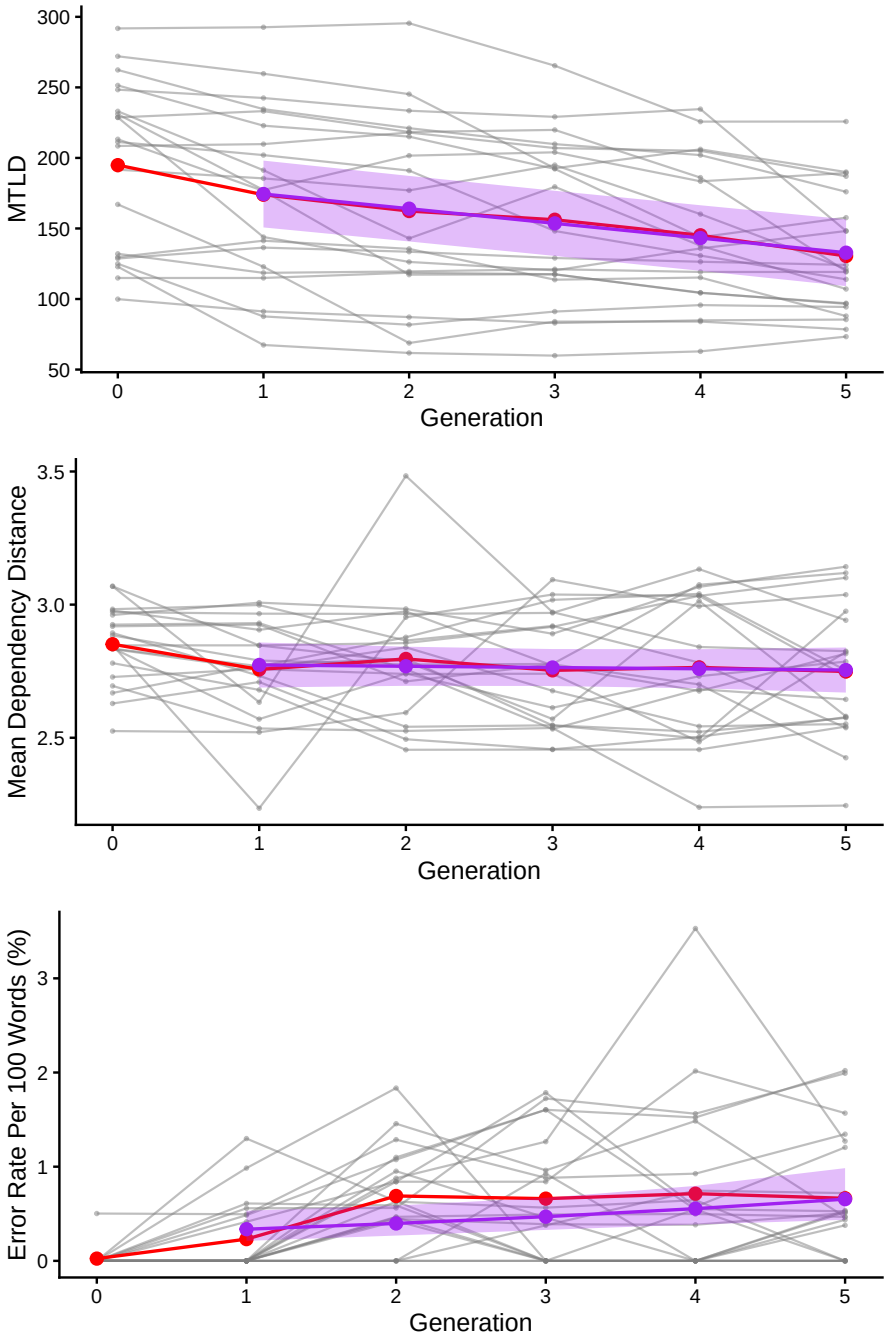


Figure 4.5: Observed and fitted trajectories of MTLD (text length-aware type-token ratio; top panel), mean dependency distance (average syntactic distance between dependent words; middle panel), and error rate (error count relative to word count; bottom panel) across generations.
In each panel, grey points show observed text-level values, and grey lines connect within-chain trajectories across generations. Red points and the red line show the mean per generation across all chains. Purple points and the purple line show model-fitted values based on mixed-effects models with random intercepts for chains (accounting for baseline variation across chains); the purple shaded bands represent 95% confidence intervals for the fixed effects. Generation 0 represents the seed texts (generated by GPT-5.2) and is shown for context; these texts were not included in the statistical models because they were generated by a different process than generations 1–5, which were created through participant edits.
MTLD shows a downward trend, indicating a decline in lexical diversity across generations. Mean dependency distance remains relatively stable across generations. Error rate shows a small increase from Generation 0 to Generation 1, followed by a relatively flat trajectory across generations 2–5, indicating that texts maintained the introduced small error rates.

and perceived effect of LLMs on one's writing style), while retaining the same random-intercept structure for chain.

In the corresponding model for MTLD, the effect of generation remained stable ($b = -10.01$, $SE = 1.46$, $t(69) = -6.83$, $p < .001$), indicating that the observed decline in MTLD is not attributable to differences in age, frequency of LLM usage, its perceived impact on society and on one's writing style. None of the additional covariates showed reliable associations with MTLD, except for two categories of perceived societal impact of LLMs. Participants endorsing "LLM chatbots will have more positive than negative effects on society" ($b = 26.83$, $SE = 11.29$, $t = 2.38$, $p = .020$) and "LLM chatbots won't have any impact on society" ($b = 38.47$, $SE = 17.63$, $t = 2.18$, $p = .032$) exhibited higher MTLD to those with no opinion. However, these seem to be isolated associations as they do not imply any meaningful pattern across the categories of perceived societal impact, and no other covariates showed significant effects. Thus, the observed decline in lexical diversity across generations cannot be attributed to individual differences in the included demographic and attitudinal variables, and these variables also do not show consistent associations with MTLD overall.

For mean dependency distance, the effect of generation remained non-significant after inclusion of the same covariates ($b = -0.014$, $SE = 0.013$, $t = -1.06$, $p = .293$), consistent with the main analysis. Among the covariates, age was negatively associated with dependency distance ($b = -0.0039$, $SE = 0.0017$, $t = -2.34$, $p = .022$), indicating that older participants produced texts with shorter mean dependency distances. In addition, participants endorsing the view that "LLM chatbots will be very beneficial to society" ($b = -0.273$, $SE = 0.127$, $t = -2.15$, $p = .035$) or "LLM chatbots will have more negative than positive effects on society" ($b = -0.230$, $SE = 0.109$, $t = -2.11$, $p = .038$) showed lower dependency distance than those with no opinion. All LLM usage frequency categories and writing-style perception variables showed no reliable associations with dependency distance.

For error count, modeled with a Poisson mixed-effects model including a log word count offset, the positive effect of generation was attenuated and no longer conventionally significant after adjustment for the same covariates ($b = 0.129$, $SE = 0.076$, $z = 1.69$, $p = .090$). None of the additional covariates showed reliable associations with error count ($p > .05$ for all LLM usage frequency categories, age, LLM impact attitudes, and writing-style perception variables).

In sum, the covariate-adjusted analyses did not alter the substantive conclusions of the main models: lexical diversity decreased robustly across generations, whereas dependency distance showed no generational change and evidence for increasing error rate remained weak. The few covariate effects that emerged (age, two attitudinal categories) are feature-specific and isolated.

4.3.3 Characterizing MTLD Decline and Error-Rate Patterns

Having established that the decline in lexical diversity is the most robust of the three confirmatory effects, and that it is not attributable to the demographic or attitudinal composition of the sample, we now examine this decline, alongside the more tentative error-rate trend, in greater detail. The following subsections ask how the MTLD decline unfolds over the course of a chain, whether it is also visible at the level of each generation's pooled output rather than only in individual texts, and what kinds of errors specifically contribute to the (non-robust) increase in error rate.

Trajectory of the Decline in Lexical Diversity

To further characterize the dynamics of the MTLD decline, we examined change relative to the immediately preceding generation (Figure 4.6a) and relative to the seed text (Figure 4.6b). MTLD differences between adjacent generations were generally negative, with an early downward shift followed by continued, though more variable, reductions across later generations. Cumulative distance from the seed text also became progressively more negative over generations, indicating that the stepwise reductions accumulated over time. This pattern suggests that the observed decline in MTLD was not limited to a few punctual early shifts, but reflected ongoing lexical simplification across iterations.

Population-Level Decline in Lexical Diversity

A decline in mean individual-text MTLD does not necessarily imply a decline in generation-level lexical diversity. In principle, reduced diversity within individual texts could be offset by complementary lexical choices across texts, such that the pooled output of a generation remains similarly diverse, or even increases in diversity overall. To assess this possibility, we additionally computed MTLD on texts concatenated within each generation. The resulting descriptive trajectory likewise declined across generations (see Figure 4.7), indicating that the reduction in lexical diversity was visible not only at the level of individual texts but also in the pooled linguistic output of each generation.

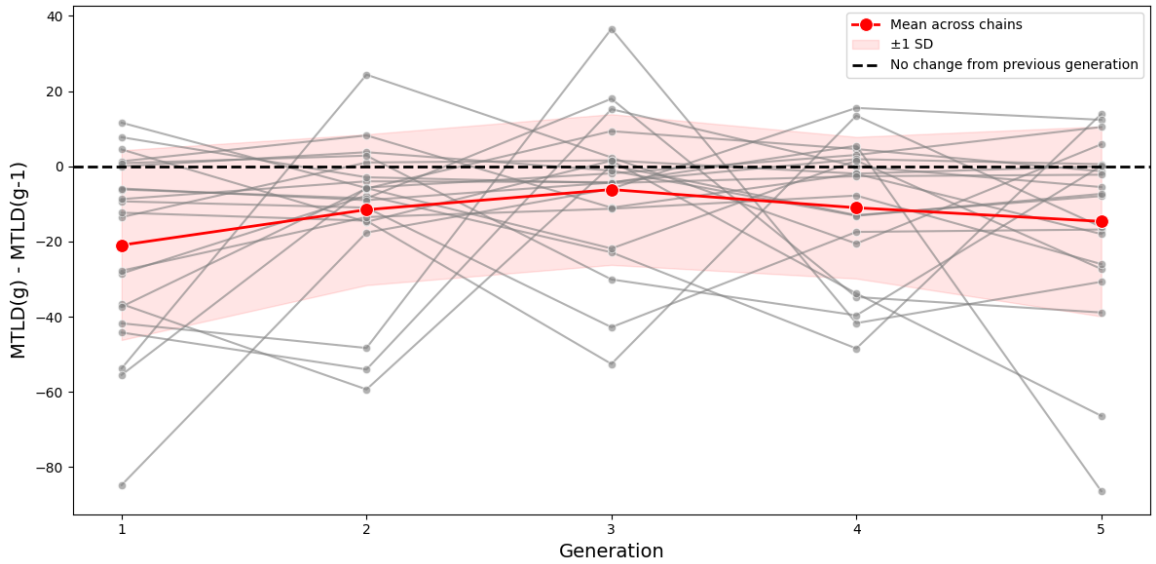
Types of Errors

To better understand the observed, though non-significant, increase in overall error rate, we also examined the types of errors that were introduced. The distribution of these error types across generations is shown in Figure 4.8. In the first generation, only grammar error were introduced, but all four types of errors were observed from the second generation onward. Punctuation and casing errors were the most common after Generation 1, while typos were relatively rare across all generations. The error rate of all types showed a decrease in Generation 3, most notably for punctuation errors, but then increased again in later generations, with grammar errors showing a particularly pronounced increase in Generation 5.

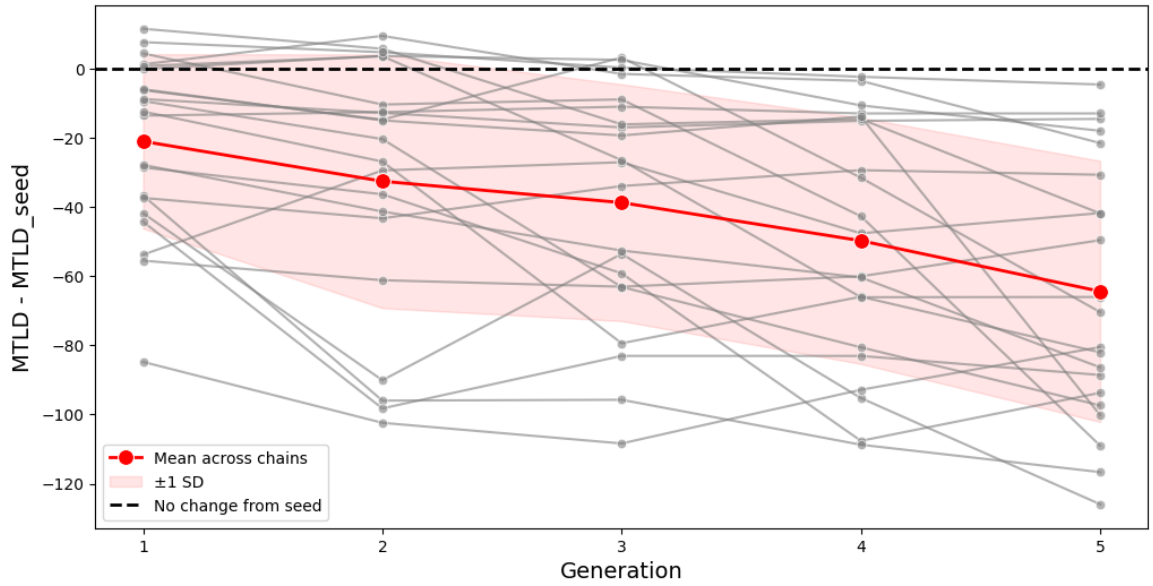
4.3.4 Sensitivity Analyses

As an additional check that the results are not a function of the model specification, we re-estimated the MTLD model using two alternative specifications. First, we fit a random-slopes model allowing the effect of generation to vary across chains, `mtld ~ generation_centered + (generation_centered | chain)`. A likelihood-ratio test indicated that this model fit significantly better than the random-intercepts model ($\chi^2(2) = 27.01$, $p < .001$). The fixed effect of generation remained negative and statistically significant ($b = -10.40$, $SE = 2.04$, $t(20) = -5.10$, $p < .001$), indicating that the decline in lexical diversity was not dependent on the assumption of a common slope across chains. The random-slopes model also indicated between-chain variation in the effect of generation on MTLD (random-slope $SD = 8.13$, variance = 66.09). The strong negative correlation between random intercepts and slopes ($r = -.82$) further suggested that chains with higher fitted MTLD levels tended to show steeper decreases across generations, whereas chains with lower fitted MTLD levels tended to decline more gradually.

Second, we fit a model with an AR(1) correlation structure to account for temporal autocorrelation



(a) MTLD difference from previous generation



(b) MTLD difference from seed text

Figure 4.6: MTLD differences between adjacent generations (left) and between each generation and the seed text (right). Each gray point represents a chain mean, with gray lines connecting chains across generations. Red points and lines represent means across chains. Red shaded areas represent 1 standard deviation from the mean. The dashed line at zero indicates no change from the previous generation (left) or from the seed text (right). The consistent negative values indicate that MTLD decreased relative to both the previous generation and the seed text across iterations.

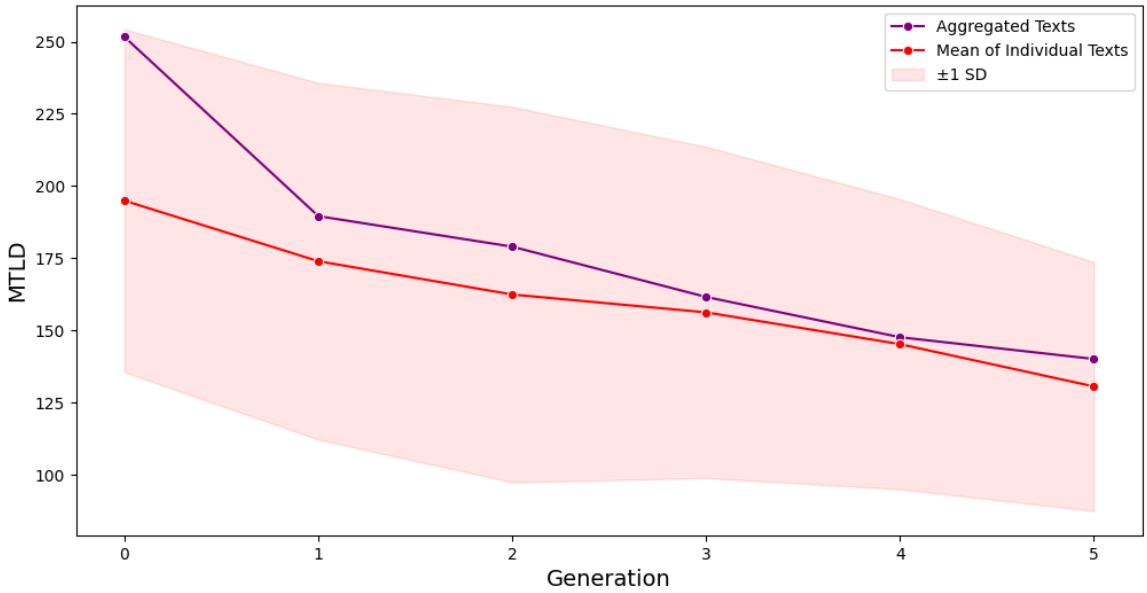


Figure 4.7: MTLD computed on concatenated texts within each generation (purple). The downward trend across generations is consistent with the decline in mean MTLD of individual texts (red), indicating that the observed reduction in lexical diversity is robust to the level of aggregation at which MTLD is computed. Generation 0 represents the seed texts

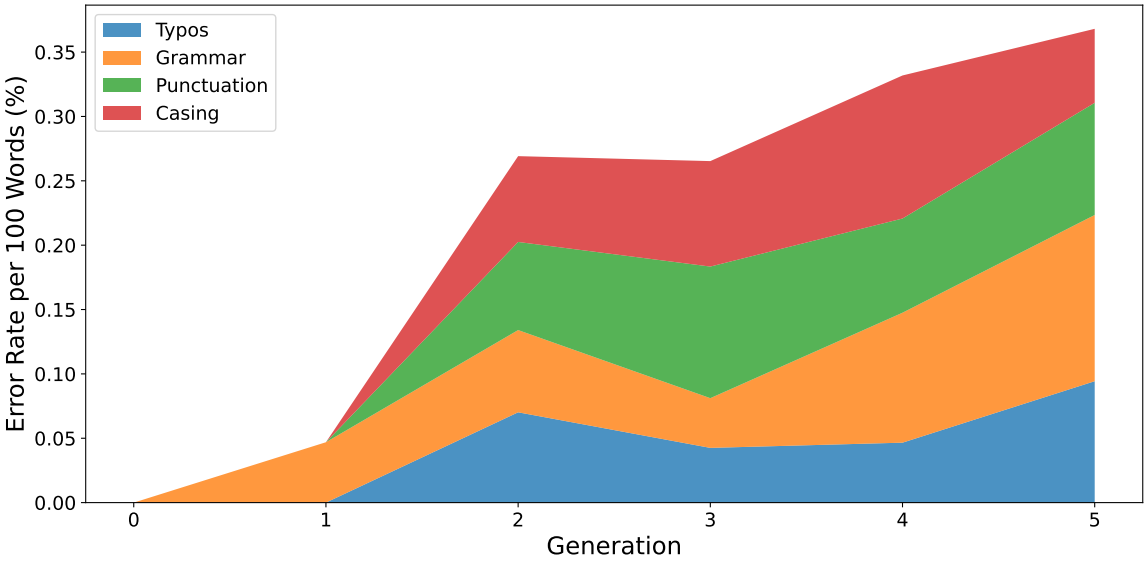


Figure 4.8: Distribution of error rates by error type across generations. The stacked areas represent the rate per 100 words of each error category (typos, grammar, punctuation, and casing). The overall height of the stacked areas at each generation corresponds to the total error rate, while the relative heights of the individual areas indicate the contribution of each error type to the total. Generation 0 represents the seed texts.
The plot shows that the error rate at Generation 1 consisted entirely of grammar errors while all later generations included all four error types, with punctuation and casing errors being most prevalent, while typos remaining relatively infrequent throughout.

among repeated observations within chains, following recommendations for analyzing iterated learning data (Winter & Wieling, 2016). This model yielded a similar estimate for generation ($b = -10.84$, $SE = 2.15$, $z = -5.05$, $p < .001$), again supporting the negative association between generation and MTLD. The estimated autocorrelation parameter was high ($r = .92$), although the residual variance was close to zero, so this model is interpreted primarily as a sensitivity analysis rather than as a preferred specification. In descriptive terms, the AR(1) model improved on the random-intercepts model in AIC (981.6 vs. 1001.8), but fit slightly worse than the random-slopes model (978.8). Across all model specifications, the estimated decline in lexical diversity was therefore highly stable, remaining close to a ten-point decrease in MTLD per generation.

To summarize, of the three text features examined, only lexical diversity showed robust, generation-linked change. MTLD declined steadily and substantially across the editing process, an effect that survived correction for multiple comparisons, replicated under covariate adjustment, and held up under two alternative model specifications addressing unequal rates of decline across chains and temporal autocorrelation within them. The decline did not arise from a few isolated edits but accumulated steadily across generations, was visible both in individual texts and in each generation's pooled output. The texts became progressively less diverse and more similar to one another across chains, a pattern that aligns with the homogenization prediction rather than the tails-based diversification prediction.

By contrast, mean dependency distance showed no evidence of change at any stage of the analysis. It behaved as a feature that fluctuates around a stable average both within and across chains, rather than as a target of the differentiation process. The error rate increased numerically across generations, particularly in the first two, but this trend did not survive correction for multiple comparisons or covariate adjustment, and a closer look at error types suggested a shifting mixture of punctuation, casing, and grammar errors.

Taken together with the previous set of results regarding editing explanation, these findings indicate that the absence of systematic change in participants' self-reported editing rationales does not extend uniformly to the texts they produced. The implication of all these results is taken up in the following chapter.

Chapter 5

Discussion

This chapter discusses the findings of the study in relation to the research questions and hypotheses that motivated the project, and situates them within the literature on language change and language use for identity signaling and negotiation. The discussion is organized around the three research questions that guided the investigation: the sustainability of iterative human-authorship signaling through language adaptation, the evolution of reported editing strategies across generations, and the evolution of text features across generations. It points out the limitations of the study along the way and concludes by suggesting directions for future research.

5.1 Sustainability of the Iterative Linguistic Adaptation

RQ1 asked to what extent individuals are capable of iteratively adapting a writing style to signal human authorship and differentiate it from one associated with LLMs that itself incorporates prior adaptations. We operationalized this question in terms of the number of iterations that participants were able to complete in the text-editing task before two consecutive participants judged their edits as unsuccessful in improving human-authorship signaling, or five iterations were completed. This was based on the assumption that if two participants in a row fail to make such improvements, then the possibility for adapting the text has been exhausted and no more participants would in general be able to successfully do so in subsequent iterations. While we cannot establish from the limited number of iterations performed in the experiment whether the process could be sustained indefinitely, the fact that no chain reached the stopping criterion suggests that it is can be sustained at least over several iterations under the experimental conditions, even without the possibility of conventionalization of new linguistic features as will be further discussed in the section regarding RQ3..

Local signs of increasing difficulty One observation from the data may be relevant to the long-term sustainability of the process. Although no two consecutive participants judged their edits as unsuccessful in any chain, five individual cases of editing failures occurred. The fact that four of these five cases occurred in the last two generations, three of them in the final generation alone, may suggest that the task became more difficult as generations progressed. This would be consistent with a gradual erosion of the capacity for human-authorship signaling through language adaptation in the studied context and raises the possibility that a sixth generation would have produced at least one further failure immediately following one of the three failures already observed in Generation 5, thus triggering early chain termination under our stopping rule. Nevertheless, the low incidence of these individual cases

(about 14% of the cases in the final generation and approximately 9.5% across the last two generations combined) indicates that individuals are generally able to find ways to insinuate their human authorship as the iterations unfolded.

Interpreting local signs of difficulty through the acts-of-identity framework If the increasing difficulty of the task is indeed a real phenomenon, it would be interesting to understand why this is the case. One tempting explanation is that the language itself sets a hard limit, that there exists some finite stock of features distinguishing human from LLM-associated writing, and that later writers simply draw it down faster than it can be replenished. This framing sits poorly with the data as failures were rare overall and the cases that did occur were concentrated at the end of the studied timeframe rather than building gradually across all five generations. It also seems to underestimate the human capacity for linguistic creativity and innovation, which has been amply documented in the historical record of language change and in the contemporary emergence of new varieties and registers (e.g., Crystal, 2011; Lau et al., 2026). Le Page and Tabouret-Keller's framework, on the other hand, offers an alternative explanation that locates the difficulty in the the social and psychological preconditions for performing an identity (Walters, 1987). It identifies four constraints that affect the success of acts of identity: the ability to identify with the desired group, the ability to access the group and recognize its behavioral patterns, sufficient motivation to align with or against the target group, and the ability to modify one's own linguistic behavior. We consider each of these constraints in turn to interpret the observed pattern of local failures and to reason about the sustainability of the process in the long-term.

The first constraint, the ability to identify with the desired group, may be considered trivially fulfilled here. It presupposes that a writer perceives humans and LLMs as distinct groups of language producers, such that they can meaningfully identify with the former and against the latter. This opposition may not be salient for some individuals in their everyday writing if they are not affected by possible suspicion of LLM use and thus do not have a particular reason to think of a human group at all. However, the experiment establishes this tension by explicitly framing a given writing style as undesirable on account of its LLM origin.

The second constraint, the ability to access the group and recognize its behavioral patterns, in our context would correspond to the ability to infer what counts as human-like writing and how it differs from LLM-generated text. This may become increasingly difficult to identify in the flux of the adaptive process as the social meaning of any given feature shifts as soon as it becomes contested ground. In our experiment, it is possible that participants were not able to infer this shifting ground and to isolate the relevant cues of LLM-ese from the two example texts they were presented with at the beginning of the main task as an illustration of the state of LLM-ese at their generation. They might have instead relied on their preconceptions of LLM output based on their experience with current LLMs. In the specific cases of the four participants who indicated editing failures, these preconceptions may have been strong as they were all frequent users of LLMs (using one at least several times a week). If these consisted mainly of associations already acted upon in earlier generations in the chain, such as removing currently flagged LLM-ese words like *delve*, later writers would have had to work harder to locate adaptations that still read as novel and human-distinguishing. However, this may not necessarily be a limitation of the experimental design, as it reflects the only slow weakening of indexical meanings (Eckert, 2019). For example, although *delve* is overused less by more recent models (Yakura et al., 2025), its association with LLM-ese is still salient.

The third constraint, sufficient motivation to align with or against the target group, is particularly relevant to explain the failures and reason about the potential sustainability of the process in the long-term. The experimental scenario emphasized the undesirability of a text being flagged as LLM-generated by an external evaluator, intended to approximate the real-world pressure for self-differentiation reviewed in Chapter 2. However, the experiment did not impose any real consequence to that suspicion. It is therefore possible that some participants did not experience the hypothesized cost as sufficiently real to motivate sustained effort, which would constitute a threat to the ecological validity of the manipulation rather than evidence that human-authorship signaling becomes intrinsically harder to motivate over time. Yet, even if the social cost of suspicion weakens over time in case LLM use becomes normalized, internally motivated differentiation may persist independently of any external sanction by a desire to not share characteristics of a group one does not identify with. This motivation could therefore sustain the process even if the external cost of suspicion weakens, but this is not something that our design can adjudicate between.

The fourth and final constraint is the ability to modify one's own linguistic behavior, which is bounded by the linguistic and cognitive resources available to the individual to perform a given act of identity. We restricted recruitment to native speakers of English specifically to hold linguistic proficiency relatively constant across participants, but cognitive constraints intrinsic to the task remain: participants had limited time to complete each edit, and devising a new adaptation under time pressure is not trivial. This may also help explain two further patterns reported below, namely the prevalence of simplification as a reported strategy and the reduction in lexical diversity observed in the resulting texts, both of which are plausibly more accessible and less cognitively demanding to execute under time pressure than alternatives, such as increasing complexity and innovating.

Objective versus subjective success The subjective nature of these factors is consistent with the operationalization of the stopping criterion in terms of participants' self-reported judgments of whether their edits successfully improved the human-authorship signal of the text. However, this leaves open the possibility that some edits judged successful by their authors would not have read as such to an outside observer, or vice versa. Having just invested effort in revising a text, participants may have been inclined to judge that effort as successful regardless of its actual effect, which would bias the results toward a more optimistic estimate of the process's sustainability than an external measure might yield. Future work could complement the present design with an additional evaluation step in which an independent panel rates the human-authorship signal of each text before and after editing, allowing the writer's own sense of success to be compared against an external judgment of it. Nonetheless, this operationalization was the more appropriate one for the phenomenon under study, namely the strategic manipulation of language by an individual seeking to signal human authorship to an external audience—a process inherently grounded in the writer's own perceptions and beliefs about what constitutes an effective signal. An objective external evaluation, while a valuable extension, would speak to a different question, namely whether writers' beliefs about successful signaling track reality, and would therefore complement rather than correct the present approach.

Taken together, these considerations suggest that the limiting factor in this process is not the language itself but writers' evolving sense of which cues still function as legible signals of humanness, and their practical capacity to act on that sense under time pressure. The second research question concerned the evolution of that sense across generations, specifically whether participants' reasoning about their

edits changed as the process unfolded. The next section addresses this question, and the subsequent one addresses the evolution of the resulting texts themselves.

5.2 Evolution of Editing Strategies across Generations

RQ2 concerned whether the reasons that individuals provided to justify their edits would change across generations, which would be indicative of a shift in the strategies they used to adapt the text to signal human authorship as it evolved.

Change in explanations was tracked through the main dimensions along which they varied, obtained by applying PCA to the semantic embeddings of the free-text responses. This approach was preferred over a predefined coding scheme, which would have risked constraining participants' creativity and biasing their reasoning toward a fixed inventory of strategies. PCA instead captured the main dimensions of variation across explanations as they actually emerged. It also offered the possibility of detecting subtler shifts in how participants reasoned, for instance, whether their language became more concrete or more abstract across generations, that a category-based coding might have missed. The trade-off is that the resulting components resist straightforward qualitative interpretation. Inspection of the extremes along each component allowed identifying broad patterns in participants' strategies, but a formal systematic analysis of the raw explanations remains a natural extension of this work.

Trajectory of change in editing strategies The confirmatory analyses did not show evidence for a systematic, monotonic change in any major dimension of variation in the explanations across generations. This would be consistent with a fixed or sticky, prior belief account, on which participants entered the task with a stable sense of what LLM-ese sounds like and applied it. This would also be consistent with the persistence of old associations in individuals' perceptions of LLM-ese even after they no longer reflect the new empirical reality (Figure 2.1). The follow-up exploratory analyses, however, provided a more nuanced picture: for three of the principal components, the trajectories suggested a quadratic pattern in which explanations initially moved in one direction and then reversed course. Although this pattern should be interpreted with caution given its exploratory status, it tentatively favors a reactive account.

Surface-level versus abstract features as editing targets While the principal components (PCs), on which the analysis of change was performed, should be interpreted as broad dimensions of variation in the free-text explanations rather than as one-to-one markers of specific editing strategies, certain PCs still appear to load more heavily on lexical changes (e.g., PC4, "discourse-level reorganization vs. surface-level correction"), whereas others emphasize structural changes (e.g., PC2, "global human-likeness judgments vs. structural editing operations"). Chapter 2 motivated a directional prediction along this dimension: because more salient, consciously noticed features tend to be more available for deliberate manipulation than abstract ones (Greenhill et al., 2017), surface-level lexical adjustments were expected to be targeted earlier in the process, with subtler structural changes emerging only once the more obvious lexical cues had been exploited.

This directional prediction was not borne out. Instead of a one-time shift from surface to structural targets, PC4 showed the same quadratic pattern noted above, with participants' focus moving from surface-level correction toward more structural reorganization and partway back again. One explanation is methodological: the experimental materials were short journalistic texts produced under an assignment

framing, which constrains how far structural reorganization can be pushed before compromising intelligibility or departing from genre norms, and this ceiling may simply have prevented a clean lexical-to-structural escalation from emerging. A second, complementary, explanation follows from the account of difficulty developed in the preceding section. If what limits writers is not the salience of a feature in the abstract but their ability to keep track of which features still count as LLM-associated at a given moment, then surface and structural features should not be expected to form a fixed difficulty hierarchy that is climbed once and for all; rather, both compete for attention as differentiation targets, and which one currently feels more diagnostic can shift as the chain's own sense of "LLM-like" is renegotiated from one generation to the next. On this reading, the absence of a linear progression is not a failure to find the predicted effect so much as evidence that salience alone is an incomplete predictor, and that the legibility of a cue, how reliably it can still be read as machine-associated, matters as much as how easily it can, in principle, be named or manipulated.

Simplification as editing operation Inspection of the principal component extremes indicates a pronounced tendency toward simplification: many participants reported choosing simpler words, shortening sentences, removing illustrative examples, and occasionally introducing informal spellings or minor orthographic irregularities. Because explanations were free-text and the PCA exploratory, this inspection is illustrative rather than a systematic coding of strategies; nonetheless, the pattern aligns with informally attested observations of writers deliberately simplifying their prose to avoid sounding like an LLM (e.g., Al-Sibai, 2025; B. Hu, 2026; Showemimo, 2025), and with formal simplification as one of the lexical strategies documented in classic descriptions of anti-language formation (Halliday, 1976; Lefkowitz & Hedgcock, 2016), a connection developed further in the following section. It is also consistent with the practical constraint on identity performance discussed above: under time pressure, with no opportunity to coordinate on a shared strategy, simplification is plausibly the lowest-cost adjustment reliably available to a writer, in contrast to structural reorganization or genuine lexical innovation, both of which demand more deliberation to execute well. Crucially, the tendency toward reduction was evident both within individual edits and across generations of our chains, suggesting that simplification functioned not as a one-off adjustment but as a durable, repeatedly reached-for strategy for signaling humanness under the conditions of this task. However, a cue this easy to deploy is also, presumably, easy for an adaptive system to absorb, which would limit its long-term diagnostic value even as it remains attractive in the short term.

Strategic awareness versus passive drift In Chapter 2 we discussed the two forces that Kristiansen (2014) distinguishes as possible drivers of social language change: *density*, the mechanical, frequency-driven effect of repeated exposure, and *subjectivity*, change motivated by speakers' social evaluations and attitudes. The participants' explanations indicate that the changes observed here are properly understood as falling on the subjectivity side of this distinction. Rather than simply drifting into a different style without reflection, participants were actively reasoning about which cues of LLM-ness mattered, repeatedly naming a small, recurring set of features, overly formal wording, overly polished or repetitive phrasing, excessive use of illustrative examples, and sentence-level structure that felt too neat or mechanical, as targets for revision. This pattern of explicit naming and targeting is also consistent with the performative view of identity work developed in Chapter 2, on which speakers do not merely inherit indexical associations passively but actively and strategically deploy them to construct a desired persona (Bucholtz & Hall, 2005; Eckert, 2019). Here, the deployment is directed not at a social group

in the usual sociolinguistic sense, but at a machine interlocutor whose stylistic profile participants treat as something to be consciously diagnosed and avoided.
The specific features participants named are not simply idiosyncratic folk beliefs invented for the task, as several of them, in particular the overuse of certain words and phrases and an unusually uniform, polished tone, correspond closely to properties independently documented as genuine statistical correlates of LLM output (Juzek & Ward, 2025; Khan et al., 2025; Muñoz-Ortiz et al., 2024). This suggests that participants' strategies, while clearly the product of subjective, lay reasoning about "what LLMs sound like," substantially overlapped with the empirically reliable cues identified in Figure 2.1 rather than being limited to the kind of miscalibrated heuristics documented elsewhere in detection tasks (Jakesch et al., 2023). Whether this overlap would hold up under a more systematic coding of strategies, and whether it would persist as the simulated LLM continued to absorb prior adaptations, remain open questions, but the explanations collected here at least indicate that participants were targeting real features of LLM-ese, even if their account of which features still served as useful signals shifted, as the quadratic patterns above suggest, over the course of the task.

5.3 Evolution of Text Features across Generations

RQ3 concerned whether the linguistic properties of the resulting written text, in particular lexical diversity, syntactic complexity, and orthographic, grammatical and punctuation correctness, would change across iterations of the process. The confirmatory analyses showed that only lexical diversity, as measured by MTLD, reliably decreased across generations after correction for multiple testing. Although error rate showed a nominal increase and dependency distance showed no directional shift, the clearest and most consistent pattern across generations was a reduction in MTLD. In substantive terms, iterative transmission appears to have made the texts lexically simpler without producing a comparable change in dependency-based syntactic structure.
One interpretation of this result is that simpler wording has become enregistered as a stable cue of human authorship and remains such throughout the iterative process. Alternatively, it is possible that the observed reduction in lexical diversity reflect the fourth constraints of Le Page and Tabouret-Keller's framework, the ability to modify one's behavior. Under this account, simplification may have emerged as the path of least resistance, that is, a lower-cost, individually executable strategy. In other words, the observed reduction in lexical diversity may not reflect a strategic choice to adopt simpler wording as a stable cue of humanness, but rather a practical constraint on writers' ability to find alternative ways to express themselves under time pressure.

Lexical change but no syntactic change Finding change in the lexical feature but not in the syntactic one is consistent with literature on language change stating that lexical change occurs faster than syntactic change and faster in features that are more salient and consciously accessible (e.g., Greenhill et al., 2017; Smitterberg, 2021). This asymmetry was reflected in participants' explanations: lexical adjustments were frequently referenced and described in concrete terms, such as replacing formal or uncommon wording with simpler alternatives, whereas references to syntactic complexity were scarce, and none of those inspected explicitly mentioned dependency structure, suggesting that this feature operated below the threshold of conscious awareness during the adaptation process. This also aligns with Walters (1987)'s fourth constraint on acts of identity, namely the ability to modify one's behaviour is bounded by what one can perceive and access. Syntactic reorganization may simply fall outside the

repertoire of adjustments writers can readily deploy.

Anti-LLM-ese as an untraditional anti-language At the same time, the direction of lexical change observed here only partially aligns with typical anti-language dynamics. While formal simplification is attested as one feature of anti-languages (Halliday, 1976; Lefkowitz & Hedgcock, 2016), it typically co-occurs with the quintessential relexicalization and lexical innovation—the continuous introduction of novel forms to maintain a counter-reality in opposition to a mainstream or out-group. In our data, simplification was present but innovation was not, suggesting an only partial instantiation of the traditional anti-language dynamic. The "brain rot" slang analyzed in Lau et al. (2026) offers a more complete instantiation. It shows that contextually grounded, fast-changing, and deliberately absurd expressions function as a social creativity strategy through which an in-group resists assimilation into AI-shaped communication by producing forms that AI struggles to replicate. The key condition enabling this process is conventionalization through repeated circulation within a community. The phenomenon of interested in our study precluded this self-differentation is an individual endeavor that does not allow communication between language producers and receivers, leaving no mechanism for novel lexical choices to spread, stabilize, or accrete into shared identity markers. Under these conditions, simplification may have emerged as the path of least resistance—a lower-cost, individually executable strategy—while the generative relexicalization more central to anti-language dynamics remained out of reach.

That said, the absence of lexical innovation does not necessarily disqualify the language emerging from our phenomenon from bearing a family resemblance to anti-language. Montgomery (2008) (cited by Lefkowitz and Hedgcock (2016)) cautions against treating anti-language as a binary category, arguing instead that particular varieties approximate its defining features to varying degrees. For instance rap culture is not separated from mainstream enough to be classified as anti-society, but nevertheless exhibits markers of an anti-language. Similarly, Lefkowitz and Hedgcock (2016) show that attested anti-languages differ considerably in which characteristics they instantiate with Verlan, for instance, showing all 15 of the identified features of an anti-language, while Heritage Spanish (Spanglish) having only nine. On this more gradient view, the language emerging from iterative human-authorship signaling may represent a new and idiosyncratic approximation of the anti-language prototype, one shaped by a boundary drawn between humans and machines rather than social groups, and one in which self-differentiation is pursued individually rather than through the collective construction of shared and secret norms. As such, it would be expected to instantiate only a subset of canonical anti-language features, with those dependent on community-level conventionalization, such as relexicalization and lexical innovation, least likely to emerge.

Possible concerns Consequently, the adversarial linguistic adaptation process does not dissipate concerns about linguistic homogenization associated with LLMs (e.g., Vanmassenhove, 2025; Yakura et al., 2025). On the contrary, it suggests that attempts to signal human authorship through surface-level stylistic adjustment may themselves push writing toward a narrower and more conventional register. Relying on the "tails" of language (Vanmassenhove, 2025), i.e., the rare, distinctive, and idiosyncratic features that often carry social meaning and individual identity, as signals of humanness may paradoxically lead to their pruning as writers gravitate toward safer, more prototypical choices. Instead, it aligns with concerns that detector-driven writing cultures encourage strategic imitation of safe patterns rather than genuinely distinctive expression (B. Hu, 2026).

Another unsettling reaction to LLM-ese that is starting to be observed is that framing humanness as imperfection, manifesting, for instance, in the deliberate introduction of errors and "dumbing down" of the prose, which is feared to lead to an intentional lowering of writing standards (Al-Sibai, 2025; B. Hu, 2026). We found some evidence of such a reaction in our data in the form of a nominal increase in error rate across generations. Participants' explanations showed that at least some of the errors were introduced intentionally.

This effect followed a different trajectory than the sustained decline in lexical diversity, namely an early increase followed by stabilization rather than a monotonic trend. This suggests that different features may be amenable to manipulation following different dynamics. The stabilization of error rate may partly reflect the constraints imposed by the experimental setting. The journalistic writing produced for an assignment imposes norms of intelligibility and formal correctness that limit how far deliberate deviation from this norm can be pushed before becoming counterproductive.

However, it is important to not to frame this variation as straightforward degradation. Usage-based and historical accounts treat language change as a normal property of linguistic systems rather than a decline in quality (e.g., Bybee, 2015), and work on internet and youth language similarly cautions against prescriptivist readings of nonstandard written registers. Tagliamonte (2016), for instance, found that internet language across email, instant messaging, and texting exhibited stable and systematic grammatical patterns rather than random deviation, while Crystal (2011) argues for treating internet language as a legitimate and rule-governed variety in its own right. At the same time, these online, informal settings, in particular synchronous text chat can blur the written-versus-spoken distinction, which may make some nonstandard or playful strategies more viable than in journalistic prose. In sum, the kinds of adaptations that can function as human-authorship signals, and the ease with which they can be deployed, seem to vary substantially across social and formal contexts, as well as across linguistic dimensions.

5.4 Future Directions

As the phenomenon under investigation, if real, is only in its early stages, the experimental design was kept relatively flexible and open-ended to allow for the emergence of unexpected patterns and to avoid constraining participants' behavior in ways that might prevent the phenomenon from manifesting. This gives space for a variety of exploratory analyses to be conducted on the obtained highly dimensional data, which are far from exhausted in this thesis.

Beyond further analysis of the existing data, the findings motivate several extensions of the design itself. The sustainability of the iterative counter-adaptation process having been demonstrated for five generations, a natural next step would be to study the phenomenon over a longer horizon. The absence of chain termination across five generations leaves open whether the capacity for iterative counter-adaptation would eventually be exhausted with more iterations. The indications of a quadratic effect along certain dimensions of editing explanations could possibly be the first part of a continuous oscillatory reaction–counterreaction dynamic in how individuals reason about their edits across generations. Furthermore, the steady decline in lexical diversity invites the question of whether this trend would continue, plateau as the space for simplification is exhausted, or eventually reverse.

To better understand the resulting dynamics, a strengthened design could include a control condition with a static rather than adaptive simulated LLM, and a parallel condition in which participants edit

without explicit differentiation instructions, to disentangle the contributions of the iterative structure and the cover scenario respectively.

The influence of LLMs on language is a topic with cross-linguistic relevance, yet research on it is still heavily English-centered. Extending this line of research beyond English would be valuable from a descriptive perspective, in keeping with the emphasis on cross-linguistic comparison in the language change literature. Such comparisons would help determine whether the patterns identified here are language- and culture-specific or instead reflect more general pressures to distinguish human-authored from machine-generated text.

Cross-linguistic work is also important from a more applied perspective. Because the training and alignment of LLMs rely in part on human feedback (Juzek & Ward, 2024), the stylistic norms that come to characterize LLM output may reflect some varieties more strongly than others. If perceptions of what counts as "LLM-like" language are shaped by these norms, speakers whose habitual writing overlaps more closely with them may face greater pressure to adapt their style in order to signal human authorship. Investigating the cross-linguistic reach of LLM-motivated linguistic change is therefore not only a descriptive question but also one with implications for linguistic equity.

Chapter 6

Conclusion

This thesis set out from an observation that, until recently, had mostly the status of an anecdote. While people grow more attuned to the linguistic fingerprints of LLM output, some appear to be deliberately reshaping their own writing to keep it from resembling LLM-generated text (Al-Sibai, 2025; AskingAboutChatGPT, 2026; B. Hu, 2026; Ice_Efficient, 2026; Si1Fei1, 2025). Given that (i) suspicion of LLM use can carry real social and institutional costs, that (ii) no detector, human or automatic, can reliably settle the question of authorship (Jakesch et al., 2023; Májovský et al., 2024), and that (iii) language is a resource through which speakers construct and signal identity in opposition to an out-group (Bucholtz & Hall, 2005; Eckert, 2019; Walters, 1987), we hypothesized that such differentiation reflects a genuine, if still emerging, process of identity signaling against LLMs. Because LLMs continue to be developed from input that may increasingly contain these very semiotic adaptations, however, we further hypothesized that any single adaptation would be short-lived, defined only in opposition to the current, folk-perceived state of LLM output, requiring the process as a whole to unfold iteratively rather than resolve in any one round of adjustment.

As this phenomenon, if real, is only in its early stages, direct observational evidence for it is limited, and the iterative dynamic in particular is not yet observable in the wild. We therefore turned to experimentation, repurposing the chain-generation structure of Iterated Learning designs (Beckner et al., 2017; Kirby, 2001; Smith et al., 2003) to simulate it under controlled conditions. We were particularly interested in how sustainable such a process could be, given that, unlike the identity-signaling dynamics typically described in the sociolinguistic literature, it affords no opportunity for individuals to communicate and conventionalize new in-group norms of humanness, and given that may unfold within the constraints of an already codified, formal writing register.

To this end, we designed an iterative editing task in which participants edited a text to make it read more convincingly as human-authored and unlike the output of a simulated, state-of-the-art LLM that had itself incorporated the adaptations of previous participants. Each participant's output was passed on as the next participant's input, organizing responses into chains of up to five generations. Alongside the edited texts, we collected participants' free-text explanations of their editing rationale and their own assessment of whether their edits had successfully strengthened the text's human-authorship signal. The latter served as our operationalization of sustainability: a chain was considered to have exhausted its capacity for further adaptation, and thus terminated early, if two consecutive participants judged their edits unsuccessful. The explanations served as a metacognitive proxy for the strategies participants drew on, while the edited texts themselves allowed us to ask whether those strategies left systematic, measurable traces in the language produced, independently of what participants reported doing.

Across these three research questions, the hypothesized process found some support, though not uniformly so. No chain terminated early, indicating that participants were able, according to their own judgement, to keep finding ways to signal human authorship across repeated rounds, though signs of growing difficulty in later generations suggest this sustainability might not be unconditional. The strategies participants reported showed no reliable evidence of systematic change across generations, despite a recurring qualitative pull toward simplification, plainer words, shorter sentences, and less illustrative detail, in participants' own descriptions of their edits. This qualitative impression was borne out at the level of the texts themselves, as lexical diversity declined steadily and robustly across generations, the only one of the three linguistic features we examined to survive correction for multiple comparisons, while syntactic complexity and error rate showed no comparable systematic change.

Taken together, these results suggest that the capacity to sustain this kind of differentiation depends jointly on what individuals believe characterizes LLM output and on their motivation to depart from it, echoing existing accounts of identity signaling as bounded by what a writer can perceive, access, and is motivated to act on (Walters, 1987). It is, at the same time, bounded by the genre and register of the chosen materials. It is possible that the formal, fact-based news article style constrained how far participants were willing to push deviations such as introducing errors before these became counterproductive, and the kinds of adaptations that prove viable here may not generalize to more informal or less codified registers.

A further constraint is the absence of any mechanism for participants to communicate and jointly negotiate what should count as a marker of humanness. While in other contexts this condition tends to give rise to genuine lexical innovation (Lau et al., 2026), its absence here appears to have funneled individual efforts toward the lowest-cost, individually achievable strategy available, namely simplification, rather than toward more inventive or idiosyncratic alternatives. When pursued widely enough, this could mean that the very act of trying to sound distinctly human nudges writing toward a narrower, more uniform register, even as each individual writer experiences their own edits as an assertion of identity rather than as conformity—the identity, however, being human nature.

We see this tension as the central message to take from this thesis, but not as cause for resignation. That participants could be observed inventing new ways to keep signaling their humanness across five successive rounds, with no clear sign of running out, is itself an indication that language remains an open and reshapable resource rather than a fixed inventory writers can only draw down. The particular adaptations observed here, narrowing rather than expanding the language used, are a product of the specific constraints of an individual, uncoordinated, and time-limited task, not of any inherent limit on what humans can do with language when distinguishing themselves from a machine. Under different conditions, longer horizons, the ability to coordinate, or less codified registers, the same impulse might just as plausibly produce invention as reduction. Our languages remain, in the end, an open and ever-expandable space in which we get to keep deciding, again and again, what it means to sound like ourselves.

Bibliography

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746 (cited on page 11).

AI writing detection: Red flags. (no date). Retrieved January 23, 2026, from https://www.montclair.edu/faculty-excellence/ofe-teaching-principles/clear-course-design/practical-responses-to-chat-gpt/red-flags-detecting-ai-writing/. (Cited on page 21).

Aitchison, J. (2013). *Language change : Progress or decay?* (4th ed.). Cambridge University Press. (Cited on pages 8, 9).

Al-Sibai, N. (2025). College Students Are Sprinkling Typos Into Their AI Papers on Purpose. Retrieved January 23, 2026, from https://futurism.com/college-students-ai-typos. (Cited on pages 5, 21, 27, 38, 69, 72, 74).

Artemova, E., Lucas, J., Venkatraman, S., Lee, J., Tilga, S., Uchendu, A., & Mikhailov, V. (2025). Beemo: Benchmark of Expert-edited Machine-generated Outputs. https://doi.org/10.48550/arXiv.2411.04032. (Cited on page 17).

AskingAboutChatGPT. (2026). AI ruined the word "delve" for me. Retrieved June 7, 2026, from https://www.reddit.com/r/antiai/comments/1snx0ib/ai%5C%5C_ruined%5C%5C_the%5C%5C_word%5C%5C_delve%5C%5C_for%5C%5C_me/. (Cited on pages 3, 20, 74).

Bao, T., Zhao, Y., Mao, J., & Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: A study on preprint papers. *Scientometrics*, *130*(7), 3597–3627. https://doi.org/10.1007/s11192-025-05341-y (cited on page 12).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01 (cited on page 37).

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176. https://doi.org/10.1093/jole/lzx001 (cited on pages 22, 25, 32, 74).

Berriche, L., & Larabi-Marie-Sainte, S. (2024). Unveiling ChatGPT text using writing style. *Heliyon*, *10*(12), e32976. https://doi.org/10.1016/j.heliyon.2024.e32976 (cited on pages 31, 34).

Bick, A., Blandin, A., & Deming, D. J. (2026). The Rapid Adoption of Generative AI. *Management Science (INFORMS)*, 1. https://doi.org/10.1287/mnsc.2025.02523 (cited on page 30).

Biever, C. (2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689. https://doi.org/10.1038/d41586-023-02361-7 (cited on page 14).

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020). Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00073 (cited on page 8).

Block, N. (1990). The Mind as the Software of the Brain. In D. N. Osherson & E. E. Smith (Editors), *An Invitation to Cognitive Science: Visual cognition. 2* (Pages 377–425). MIT Press. (Cited on page 14).

Bloomfield, L. (1984). *Language* (C. F. Hackett, Editor). University of Chicago Press. Retrieved June 5, 2026, from https://press.uchicago.edu/ucp/books/book/chicago/L/bo3636364.html. (Cited on page 19).

Bochkarev, V., Solovyev, V., & Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, *11*(101), 20140841. https://doi.org/10.1098/rsif.2014.0841 (cited on pages 8, 9, 18, 21).

Boutadjine, A., Harrag, F., & Shaalan, K. (2025). Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *24*(2), 12:1–12:26. https://doi.org/10.1145/3708889 (cited on page 15).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165. (Cited on page 10).

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, *7*(4-5), 585–614. https://doi.org/10.1177/1461445605054407 (cited on pages 4, 19, 69, 74).

Bybee, J. L. (2015). *Language change*. Cambridge University Press. (Cited on pages 7, 8, 72).

Cabanac, G., Labbé, C., & Magazinov, A. (2021). Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. https://doi.org/10.48550/arXiv.2107.06751. (Cited on page 16).

Cai, S., & Cui, W. (2023). Evade ChatGPT Detectors via A Single Space. Retrieved June 2, 2026, from https://arxiv.org/abs/2307.02599v2. (Cited on page 17).

Cardon, P. W., & Coman, A. W. (2025). Professionalism and Trustworthiness in AI-Assisted Workplace Writing: The Benefits and Drawbacks of Writing With AI. *International Journal of Business Communication*, 23294884251350599. https://doi.org/10.1177/23294884251350599 (cited on page 17).

Chaib, O. (2026). Tropes - AI Writing Pattern Directory. Retrieved June 5, 2026, from https://tropes.fyi/. (Cited on pages 3, 13, 21).

Chakraborty, M., Tonmoy, S. T. I., Zaman, S. M. M., Gautam, S., Kumar, T., Sharma, K., Barman, N., Gupta, C., Jain, V., Chadha, A., Sheth, A., & Das, A. (2023). Counter Turing Test (CT2): AI-Generated Text Detection is Not as Easy as You May Think - Introducing AI Detectability Index (ADI). In H. Bouamor, J. Pino, & K. Bali (Editors), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Pages 2206–2239). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.136. (Cited on page 14).

Cheng, A., Calhoun, A., & Reedy, G. (2025). Artificial intelligence-assisted academic writing: Recommendations for ethical use. *Advances in Simulation*, *10*, 22. https://doi.org/10.1186/s41077-025-00350-6 (cited on page 16).

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger. (Cited on page 6).

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. https://doi.org/10.48550/arXiv.2107.00061. (Cited on page 15).

Cornish, H., Dale, R., Kirby, S., & Christiansen, M. H. (2017). Sequence Memory Constraints Give Rise to Language-Like Structure through Iterated Learning. *PLOS ONE*, *12*(1), e0168532. https://doi.org/10.1371/journal.pone.0168532 (cited on pages 22, 32).

Creo, A., & Pudasaini, S. (2025). SilverSpeak: Evading AI-Generated Text Detectors using Homoglyphs. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, & G. Mikros (Editors), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)* (Pages 1–46). International Conference on Computational Linguistics. Retrieved February 6, 2026, from https://aclanthology.org/2025.genaidetect-1.1/. (Cited on pages 17, 27).

Croft, W. (2000). *Explaining language change : An evolutionary approach*. Longman. (Cited on pages 7, 19).

Crystal, D. (2011). *Internet linguistics : A student guide* (null, Volume null). Routledge. https://research.ebsco.com/linkprocessor/plink?id=ca3b5954-c4aa-3722-81f7-452e925eadf0. (Cited on pages 9, 66, 72).

Darda, K. M., PhD, Carré, M., & Cross, E. S. (2022). Original or fake? Value attributed to text-based archives generated by artificial intelligence. https://doi.org/10.31234/osf.io/s92am. (Cited on page 18).

David, I., & Gervais, A. (2025). AuthorMist: Evading AI Text Detectors with Reinforcement Learning. https://doi.org/10.48550/arXiv.2503.08716. (Cited on page 17).

Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, & S. Szpakowicz (Editors), *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (Pages 22–33). Association for Computational Linguistics. Retrieved March 9, 2026, from https://aclanthology.org/W18-4503/. (Cited on pages 8, 9, 12, 18, 21).

Doru, B., Maier, C., Busse, J. S., Lücke, T., Schönhoff, J., Enax- Krumova, E., Hessler, S., Berger, M., & Tokic, M. (2025). Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study. *JMIR Medical Education*, *11*, e62779. https://doi.org/10.2196/62779 (cited on page 16).

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, *41*(Volume 41, 2012), 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828 (cited on page 18).

Eckert, P. (2019). The Limits of Meaning: Social Indexicality, Variation, and the Cline of Interiority. *Language*, *95*(4), 751–776. Retrieved March 13, 2026, from https://www.jstor.org/stable/48771170 (cited on pages 4, 18, 66, 69, 74).

Fajt, B., & Schiller, E. (2025). ChatGPT in Academia: University Students' Attitudes Towards the use of ChatGPT and Plagiarism. *Journal of Academic Ethics*, *23*(3), 1363–1382. https://doi.org/10.1007/s10805-025-09603-5 (cited on page 16).

Fariello, S., Fenza, G., Forte, F., Gallo, M., & Marotta, M. (2025). Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, *9*(3), 6–18. https://doi.org/10.9781/ijimai.2024.12.002 (cited on page 16).

Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, *49*, 100321. https://doi.org/10.1016/j.iree.2025.100321 (cited on pages 15, 16, 31, 34).

Fink, C. (2025). Seven Deadly Tells Of AI Writing. Retrieved June 5, 2026, from https://www.forbes.com/sites/charliefink/2025/06/12/the-seven-tells-of-ai-writing/. (Cited on pages 3, 13, 21).

Fraud Detection - Qualtrics. (no date). Retrieved May 28, 2026, from https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/. (Cited on page 34).

Freeman, J. (2025). *Student Generative AI Survey 2025* (Policy Note Number 61). (Cited on page 15).

Galpin, R., Anderson, B., & Juzek, T. S. (2025). Exploring the Structure of AI-Induced Language Change in Scientific English. *The International FLAIRS Conference Proceedings*, *38*. https://doi.org/10.32473/flairs.38.1.138958 (cited on pages 11, 12).

Geng, M., Chen, C., Wu, Y., Wan, Y., Zhou, P., & Chen, D. (2025). The Impact of Large Language Models in Academia: From Writing to Speaking. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Editors), *Findings of the Association for Computational Linguistics: ACL 2025* (Pages 19303–19319). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.987. (Cited on pages 11, 13).

Geng, M., & Trotta, R. (2025). Human-LLM Coevolution: Evidence from Academic Writing. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Editors), *Findings of the Association for Computational Linguistics: ACL 2025* (Pages 12689–12696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.657. (Cited on pages 3, 5, 11, 20, 27).

Georgiou, G. P. (2026). Key Features to Distinguish Between Human- and AI-Generated Texts: Perspectives from University Professors. *AI in Education*, *2*(1). https://doi.org/10.3390/aieduc2010002 (cited on pages 15, 27, 38).

Goodchild, L., Mulligan, A., Calero, M. A., & Mansell, N. (2024). *Insights 2024: Attitudes toward AI* (technical report). Elsevier. Retrieved April 21, 2026, from https://www.elsevier.com/insights/attitudes-toward-ai. (Cited on pages 12, 15).

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), E8822–E8829. https://doi.org/10.1073/pnas.1700388114 (cited on pages 4, 8, 9, 22, 23, 68, 70).

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., Murakami, A., Perlman, M., Roemling, D., & Winter, B. (2025). The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1472411 (cited on page 7).

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In D. Jurafsky,

J. Chai, N. Schluter, & J. Tetreault (Editors), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Pages 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740. (Cited on page 10).

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. In M. Lapata & H. T. Ng (Editors), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Pages 363–371). Association for Computational Linguistics. Retrieved April 9, 2026, from https://aclanthology.org/D08-1038/. (Cited on page 8).

Halliday, M. a. K. (1976). Anti-Languages. *American Anthropologist*, *78*(3), 570–584. https://doi.org/10.1525/aa.1976.78.3.02a00050 (cited on pages 19, 20, 69, 71).

Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (no date). News-please: A Generic News Crawler and Extractor. *Proceedings of the 15th International Symposium of Information Science*, 218–223. Retrieved March 24, 2026, from https://huggingface.co/datasets/vblagoje/cc%5C%5C_news/viewer/plain%5C%5C_text/train (cited on page 31).

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022 (cited on page 13).

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. https://doi.org/10.48550/arXiv.2401.12070. (Cited on page 17).

Hansen, L., Olsen, L. R., & Enevoldsen, K. (2023). TextDescriptives: A Python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, *8*(84), 5153. https://doi.org/10.21105/joss.05153 (cited on page 38).

Harris, S. (2026). The Bittersweet Age. Retrieved June 3, 2026, from https://www.samharris.org/podcasts/making-sense-episodes/476-the-bittersweet-age. (Cited on page 13).

Henestrosa, A. L., & Kimmerle, J. (2024). The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text. *Journalism and Media*, *5*(3), 1085–1097. https://doi.org/10.3390/journalmedia5030069 (cited on page 18).

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9 (cited on pages 31, 34, 38).

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 5487. https://doi.org/10.1038/s41598-023-30938-9 (cited on page 17).

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. Retrieved May 29, 2026, from https://www.jstor.org/stable/4615733 (cited on page 40).

Hopper, P. J. (2003). *Grammaticalization* (2nd ed). Cambridge University Press. (Cited on page 8).

How Prolific detects bots and AI in online research. (no date). Retrieved May 28, 2026, from https://www.prolific.com/resources/how-prolific-detects-bots-and-ai-in-online-research. (Cited on page 34).

Hu. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Retrieved February 6, 2026, from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (cited on page 3).

Hu, B. (2026). Why artificial intelligence detectors could penalize academic writing. *Nature Human Behaviour*, 1–1. https://doi.org/10.1038/s41562-026-02420-9 (cited on pages 5, 13, 21, 23, 27, 38, 69, 71, 72, 74).

Ice_Efficient. (2026). It makes me so mad that I can't use an em dash normally anymore. Retrieved June 7, 2026, from https://www.reddit.com/r/RandomThoughts/comments/1tlyiov/it%5C%5C_makes%5C%5C_me%5C%5C_so%5C%5C_mad%5C%5C_that%5C%5C_i%5C%5C_cant%5C%5C_use%5C%5C_an%5C%5C_em%5C%5C_dash/. (Cited on pages 3, 20, 74).

Ireland, M. E., & Mehl, M. R. (2014). Natural Language Use as a Marker of Personality. In T. M. Holtgraves (Editor), *The Oxford Handbook of Language and Social Psychology* (Pages 201–218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199838639.013.034. (Cited on page 18).

Jack Morris. (2026). Language_tool_python: Checks grammar using LanguageTool. https://pypi.org/project/language%5C%5C_tool%5C%5C_python/. (Cited on page 39).

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300469 (cited on page 17).

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120 (cited on pages 15, 23, 27, 70, 74).

Jeremy Nguyen [@JeremyNguyenPhD]. (2024). Are medical studies being written with ChatGPT? Well, we all know ChatGPT overuses the word "delve". Look below at how often the word 'delve' is used in papers on PubMed (2023 was the first full year of ChatGPT). https://t.co/iNxZfFLkxL. Retrieved March 15, 2026, from https://x.com/JeremyNguyenPhD/status/1774021645709295840. (Cited on page 11).

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100093. https://doi.org/10.1016/j.chbah.2024.100093 (cited on page 17).

Johnstone, B. (2016). Enregisterment: How linguistic items become linked with ways of speaking. *Language and Linguistics Compass*, *10*(11), 632–643. https://doi.org/10.1111/lnc3.12210 (cited on page 19).

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed. 2002.). Springer New York. https://doi.org/10.1007/b98835. (Cited on page 36).

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd). https://web.stanford.edu/%5Ctextasciitilde%20jurafsky/slp3/. (Cited on page 10).

Juzek, T. S., & Ward, Z. B. (2024). Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. https://doi.org/10.48550/arXiv.2412.11385. (Cited on pages 3, 4, 10, 73).

Juzek, T. S., & Ward, Z. B. (2025). Word Overuse and Alignment in Large Language Models: The Influence of Learning from Human Feedback. https://doi.org/10.17605/OSF.IO/JY48S. (Cited on pages 10, 70).

Kashtan, I., & Kipnis, A. (2024). An Information-Theoretic Approach for Detecting Edits in AI-Generated Text. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.5dbf3265 (cited on page 17).

Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2025). A Survey of Reinforcement Learning from Human Feedback. https://doi.org/10.48550/arXiv.2312.14925. (Cited on page 10).

Khan, H. U., Naz, A., Alarfaj, F. K., & Almusallam, N. (2025). Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques. *Scientific Reports*, *15*(1), 20366. https://doi.org/10.1038/s41598-025-08208-7 (cited on pages 10, 70).

Kirby, S. (2001). Spontaneous evolution of linguistic structure - An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110. https://doi.org/10.1109/4235.918430 (cited on pages 22, 25, 74).

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105 (cited on pages 4, 22, 25, 32).

Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., & Goldstein, T. (2024). On the Reliability of Watermarks for Large Language Models. https://doi.org/10.48550/arXiv.2306.04634. (Cited on page 16).

Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, *186*, 114984. https://doi.org/10.1016/j.jbusres.2024.114984 (cited on page 17).

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, *11*(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813 (cited on pages 11, 21).

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, 106553. https://doi.org/10.1016/j.chb.2020.106553 (cited on page 15).

Kousha, K., & Thelwall, M. (2026). How much are LLMs changing the language of academic papers after ChatGPT? A multi-database and full text analysis. https://doi.org/10.48550/arXiv.2509.09596. (Cited on pages 3, 11, 21).

Kreps, S. E., McCain, M., & Brundage, M. (2020). All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. https://doi.org/10.2139/ssrn.3525002. (Cited on page 15).

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. https://doi.org/10.48550/arXiv.2303.13408. (Cited on pages 16, 17).

Kristiansen, T. (2014). Knowing the driving force in language change: Density or subjectivity? *Journal of Sociolinguistics*, *18*(2), 233–241. https://doi.org/10.1111/josl.12073 (cited on pages 19, 69).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13 (cited on page 37).

Labov, W. (1994). *Principles of linguistic change* (Volume 1). Blackwell. (Cited on pages 7, 9, 19).

LanguageTool. (no date). Retrieved May 28, 2026, from https://languagetool.org/. (Cited on page 38).

Lau, G. R., Koh, S. M., & Hartanto, A. (2026). Brain Rot Slang as AI Resistance: A Socio-Motivational Framework of Youth Identity in Social Media. Retrieved June 4, 2026, from https://osf.io/preprints/psyarxiv/qr3cy%5C%5C_v2/. (Cited on pages 21, 66, 71, 75).

Lefkowitz, N., & Hedgcock, J. S. (2016). 13. Anti-language: Linguistic innovation, identity construction, and group affiliation among emerging speech communities. In N. Bell (Editor), *Multiple Perspectives on Language Play* (Pages 347–376). De Gruyter Mouton. Retrieved February 6, 2026, from https://www.degruyterbrill.com/document/doi/10.1515/9781501503993-014/html. (Cited on pages 20, 69, 71).

Leiter, C., Belouadi, J., Chen, Y., Zhang, R., Larionov, D., Kostikova, A., & Eger, S. (2024). NLLG Quarterly arXiv Report 09/24: What are the most influential current AI Papers? https://doi.org/10.48550/arXiv.2412.12121. (Cited on page 20).

Lewis, D. (2008). *Convention: A Philosophical Study*. Wiley-Blackwell. (Cited on page 7).

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7). https://doi.org/10.1016/j.patter.2023.100779 (cited on page 17).

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers. https://doi.org/10.48550/arXiv.2404.01268. (Cited on page 11).

Májovský, M., Černý, M., Netuka, D., & Mikolov, T. (2024). Perfect detection of computer-generated text faces fundamental challenges. *Cell Reports Physical Science*, *5*(1), 101769. https://doi.org/10.1016/j.xcrp.2023.101769 (cited on pages 16, 17, 27, 74).

Masrour, E., Emi, B. N., & Spero, M. (2025). DAMAGE: Detecting Adversarially Modified AI Generated Text. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, & G. Mikros (Editors), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)* (Pages 120–133). International Conference on Computational Linguistics. Retrieved February 6, 2026, from https://aclanthology.org/2025.genaidetect-1.9/. (Cited on page 17).

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381 (cited on page 38).

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547 (cited on page 38).

Milička, J., Marklová, A., Drobil, O., & Pospíšilová, E. (2025). Humans can learn to detect AI-generated texts, or at least learn when they can't. *PLOS One*, *20*(10), e0333007. https://doi.org/10.1371/journal.pone.0333007 (cited on page 16).

Millar, R. M., & Trask, R. L. (2023). *Trask's Historical Linguistics* (4th edition). Routledge. https://doi.org/10.4324/9781003125136. (Cited on page 21).

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, *4*(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106 (cited on pages 38, 39).

Montgomery, M. (2008). *An introduction to language and society* (3rd edition). Routledge. (Cited on page 71).

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, *57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2 (cited on pages 10, 38, 70).

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., & Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In C. Callison-Burch & M. Dredze (Editors), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Pages 122–130). Association for Computational Linguistics. Retrieved February 12, 2026, from https://aclanthology.org/W10-0719/. (Cited on page 34).

Opara, C. (2025). Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis. https://doi.org/10.48550/arXiv.2505.01800. (Cited on pages 17, 31, 34).

OpenAI. (2022). Introducing ChatGPT. Retrieved April 5, 2026, from https://openai.com/index/chatgpt/. (Cited on page 3).

Padillah, R. (2024). Ghostwriting: A reflection of academic dishonesty in the artificial intelligence era. *Journal of Public Health*, *46*(1), e193–e194. https://doi.org/10.1093/pubmed/fdad169 (cited on page 17).

Padmakumar, V., & He, H. (2024). Does Writing with Language Models Reduce Content Diversity? https://doi.org/10.48550/arXiv.2309.05196. (Cited on page 10).

Pagel, M., Beaumont, M., Meade, A., Verkerk, A., & Calude, A. (2019). Dominant words rise to the top by positive frequency-dependent selection. *Proceedings of the National Academy of Sciences*, *116*(15), 7397–7402. https://doi.org/10.1073/pnas.1816994116 (cited on page 8).

Partington, S., Kamtekar, R., & Nichols, S. (2025). Norms emerge through iterated learning. *Proceedings of the National Academy of Sciences*, *122*(29), e2504178122. https://doi.org/10.1073/pnas.2504178122 (cited on page 32).

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A. S., Shawe-Taylor, J., & Vespignani, A. (2025). Human-AI coevolution. *Artificial Intelligence*, *339*, 104244. https://doi.org/10.1016/j.artint.2024.104244 (cited on page 13).

Prolific. (2026). Retrieved May 22, 2026, from https://www.prolific.com. (Cited on page 34). Online participant recruitment platform.

Pudasaini, S., Miralles, L., Lillis, D., & Salvador, M. L. (2025). Benchmarking AI text detection: Assessing detectors against new datasets, evasion tactics, and enhanced llms. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, & G. Mikros (Editors), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)* (Pages 68–77). International Conference on Computational Linguistics. Retrieved January 25, 2026, from https://aclanthology.org/2025.genaidetect-1.4/. (Cited on page 16).

Raj, M., Berg, J. M., & Seamans, R. (2026). The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*, *155*(4), 896–915. https://doi.org/10.1037/xge0001889 (cited on page 18).

Rathi, I., Taylor, S., Bergen, B. K., & Jones, C. R. (2024). GPT-4 is judged more human than humans in displaced and inverted Turing tests. https://doi.org/10.48550/arXiv.2407.08853. (Cited on page 15).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084. (Cited on page 36).

Russell, J., Karpinska, M., & Iyyer, M. (2025). People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Editors), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Pages 5342–5373). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.267. (Cited on pages 15, 27, 30, 38).

Sam Altman [@sama]. (2025). I have had the strangest experience reading this: I assume its all fake/bots, even though in this case i know codex growth is really strong and the trend here is real. i think there are a bunch of things going on: Real people have picked up quirks of LLM-speak, the Extremely. Retrieved April 10, 2026, from https://x.com/sama/status/1965110064215458055. (Cited on page 13).

Sandler, M., Choung, H., Ross, A., & David, P. (2025). A Linguistic Comparison Between Human and ChatGPT-Generated Conversations. In C. Wallraven, C.-L. Liu, & A. Ross (Editors), *Pattern Recognition and Artificial Intelligence* (Pages 366–380). Springer Nature. https://doi.org/10.1007/978-981-97-8702-9_25. (Cited on pages 31, 34).

Sasaki, Y. (2017). Publishing Nations: Technology Acquisition and Language Standardization for European Ethnic Groups. *The Journal of Economic History*, *77*(4), 1007–1047. https://doi.org/10.1017/S0022050717000821 (cited on page 9).

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763. https://doi.org/10.1213/ANE.0000000000002864 (cited on page 11).

Searle, J. R. (1990). Is the Brain's Mind a Computer Program? *Scientific American*, *262*(1), 25–31. Retrieved June 19, 2026, from https://www.jstor.org/stable/24996641 (cited on page 14).

Shapira, P. (2024). Delving into "delve". Retrieved March 15, 2026, from https://pshapira.net/2024/03/31/delving-into-delve/. (Cited on pages 3, 11).

Showemimo, P. (2025). The AI Paradox in Academia: When Writing Too Well Raises Red Flags. https://doi.org/10.2139/ssrn.5255976. (Cited on pages 17, 21, 69).

Si1Fei1. (2025). I feel so resentful … Retrieved June 7, 2026, from https://www.reddit.com/r/OpenAI/comments/1mk62b1/em%5C%5C_dashes%5C%5C_are%5C%5C_back%5C%5C_baby/n7gnqpb/. (Cited on pages 3, 20, 74).
Post URL: https://www.reddit.com/r/OpenAI/comments/1mk62b1/em_dashes_are_back_-baby/.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated Learning: A Framework for the Emergence of Language. *Artificial Life*, *9*(4), 371–386. https://doi.org/10.1162/106454603322694825 (cited on pages 22, 74).

Smitterberg, E. (2021). Aspects of Language Change. In *Syntactic Change in Late Modern English: Studies on Colloquialization and Densification* (Pages 42–76). Cambridge University Press. https://doi.org/10.1017/9781108564984.003. (Cited on pages 4, 7, 70).

Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., Tak, A., Chen, M., Morstatter, F., & Dehghani, M. (2025). The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models. Retrieved April 14, 2026, from https://arxiv.org/abs/2502.11266v1. (Cited on page 31).

Swadesh, M. (2006). *The origin and diversification of language*. Adeline Transaction. (Cited on page 8).

Székely, É., Miniota, J., Míša, & Hejná. (2025). Will AI shape the way we speak? The emerging sociolinguistic influence of synthetic voices. https://doi.org/10.48550/arXiv.2504.10650. (Cited on page 13).

Tagliamonte, S. A. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, *45*(1), 1–32. https://doi.org/10.1017/S0047404515000780 (cited on pages 9, 72).

Thompson, A. L., Hoey, T. V., Chik, A. W. C., & Do, Y. (2025). Iconic hand gestures from ideophones exhibit stability and emergent phonological properties: An iterated learning study. *Cognitive Linguistics*, *36*(2), 227–259. https://doi.org/10.1515/cog-2024-0033 (cited on page 32).

Tsigaris, P., & Teixeira da Silva, J. A. (2026). AI detecting AI in academic writing: Why most AI detector findings are false. *Next Research*, *7*, 101396. https://doi.org/10.1016/j.nexres.2026.101396 (cited on page 16).

Turing, A. M. (1950). I.-Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433 (cited on page 14).

Van Noorden, R., & Perkel, J. M. (2023). AI and science: What 1,600 researchers think. *Nature*, *621*(7980), 672–675. https://doi.org/10.1038/d41586-023-02980-0 (cited on pages 3, 12, 15).

Vanmassenhove, E. (2025). Losing our Tail – Again: On (Un)Natural Selection And Multilingual Large Language Models. https://doi.org/10.48550/arXiv.2507.03933. (Cited on pages 9, 23, 71).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. Retrieved January 4, 2024, from https://proceedings.neurips.cc/paper%5C%5C_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (cited on page 10).

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2024). Ghostbuster: Detecting Text Ghostwritten by Large Language Models. https://doi.org/10.48550/arXiv.2305.15047. (Cited on page 17).

Walters, K. (1987). Review of Acts of Identity: Creole-Based Approaches to Language and Ethnicity. *Language in Society*, *16*(4), 569–571. Retrieved March 31, 2026, from https://www.jstor.org/stable/4167885 (cited on pages 4, 66, 70, 74, 75).

Wang, S., & Huang, G. (2024). The Impact of Machine Authorship on News Audience Perceptions: A Meta-Analysis of Experimental Studies. *Communication Research*, *51*(7), 815–842. https://doi.org/10.1177/00936502241229794 (cited on page 17).

Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2025). History, development, and principles of large language models: An introductory survey. *AI and Ethics*, *5*(3), 1955–1971. https://doi.org/10.1007/s43681-024-00583-7 (cited on page 10).

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International*

Journal for Educational Integrity, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z (cited on pages 15, 16).

Weizenbaum, J., & View Profile. (1983). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *26*(1), 23–28. https://doi.org/10.1145/357980.357991 (cited on page 14).

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, *1*(1), 7–18. https://doi.org/10.1093/jole/lzv003 (cited on pages 39, 64).

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I., & Rahwan, I. (2025). Empirical evidence of Large Language Model's influence on human spoken communication. https://doi.org/10.48550/arXiv.2409.01754. (Cited on pages 3, 4, 11–13, 31, 34, 66, 71).

Zamaraeva, O., Flickinger, D., Bond, F., & Gómez-Rodríguez, C. (2025). Comparing LLM-generated and human-authored news text using formal syntactic theory. https://doi.org/10.48550/arXiv.2506.01407. (Cited on page 10).

Appendix A

Supplementary Materials Related to Methods

A.1 A Formalization of the Iterative Editing Task

Each inherited target text $T_{g,c}$ at generation g in chain c is a concatenation of n sentences: $T_{g,c} := \|_{i=1}^{n_{g,c}} s_{g,c,i}$. Each sentence $s_{g,c,i}$ is the result of the application of a set of edit operations[1] $E(\cdot)$ to sentence $s_{g-1,c,i}$ by a participant p_j, $j \in \{1; 105\}$. Since each text $T_{g,c}$ is inherited and edited by a unique participant, participant and position in the chain-generation structure are implementationally confounded, that is, $p_j \equiv (g, c)$, then $s_{g,c,i} = E(s_{g-1,c,i}, p)$ simplifies to $s_{g,c,i} = E(s_{g-1,c,i})$, although conceptually distinct.

Each sentence modification $E(s_{g,c,i})$ is the result/product of some underlying pattern of reasoning employed by participant (g, c). Assuming that participants attempted to meet the explicit goal of the editing task (ensured by attention checks and response coherence), namely to make $T_{g,c}$ read convincingly as human-written, we consider the observed modifications not to be arbitrary transformations but strategically guided actions directed toward that objective. We therefore construe the reasoning pattern underlying each sentence modification $E(s_{g,c,i})$ as an application $z_{g,c,i}$ of some latent human-authenticating strategy $S_k \in \mathcal{S}$, where $\mathcal{S} = \{S_1, \ldots, S_K\}$ is a set of latent human-authenticating strategies. Note that we make the simplifying assumption that a single strategy was used in the editing of each sentence. This allows using the sentence as a unit of analysis, circumventing the problem of defining what constitutes one edit arbitrarily, at the character-, word, or other level. While the actual edits applied to a sentence may suggest a type of underlying strategy, they constitute only very indirect clues. To tighten this interpretation gap, participants were also asked to provide a free-text explanation for each sentence they modified. These reports do not constitute concrete strategy labels, but can serve as proxies for the instantiated human-authenticating strategies. Let $X_{g,c,i}$ denote the participant-provided explanation associated with the modification of sentence $s_{g,c,i}$. We treat $X_{g,c,i}$ as a metacognitive report that provides indirect evidence about the instantiated strategy $z_{g,c,i}$, but that is

[1] Although individual edit operations are applied in a sequence by participants while taking the experiment, only the final sentence state is recorded. The sequential nature of the editing process is therefore collapsed into a set of differences between each inherited sentence $s_{g-1,c,i}$ and it edited version $s_{g,c,i}$

more direct than the observed text transformation $E(s_{g-1,c,i})$. Thus:

$$S_k \xrightarrow{\text{instantiates}} z_{g,c,i} \xrightarrow{\text{produces}} \begin{cases} E(s_{g-1,c,i}) \\ X_{g,c,i} \end{cases} \tag{A.1}$$

where $X_{g,c,i}$ is the explanation text.

Assuming the modification $E(s_{g,c,i})$ is strategic and oriented towards the task goal, we identify several factors on which we predict its instantiation to be dependent. First, it is conditioned on (some set of) the features of the inherited sentence $s_{g,c,i}$ to which it applies. Second, it is dependent on participant (g, c)'s preconception of human and LLM writing style, which inform the human-authenticating strategy S_k of which $E(s_{g,c,i})$ is the result. These preconceptions are themselves informed by the participant's experience with using LLMs. Thirdly, what concrete edits are made is guided by the participant's own writing style and linguistic preferences, and thus might differ between participants editing the same text and using the same human-authenticating strategy.

Let $z_{g,c,i}$ denote the instantiation of some latent strategy S_k in sentence $s_{g,c,i}$, where $\mathcal{S} = \{S_1, \ldots, S_K\}$ is a set of latent strategies. Note that two instantiations of the same strategy S_k may often lead to different resulting text states, not only because of the three aforementioned factors, but also because the same strategy may admit multiple realizations. For example, if S_k is "using simplifying language" and the sentence to which it is applied is "The methodology employed in this study yielded significant improvements", then $z_{g,c,i}$ could be any of the following: "The method used in this study worked much better," "This approach gave better results," or "The study's method improved the results a lot." Similarly, multiple observable edit realizations $E(s_{g,c,i})$ may instantiate the same latent strategy S_k, since strategies admit multiple lexical and syntactic realizations conditioned on sentence context and participant-specific linguistic preferences.

Given the above, we can conceive of a conceptual model of the text-editing transmission structure as follows:

$$T_{g-1,c} \xrightarrow{\text{elicits/conditions}} S_{g,c} \xrightarrow{\text{instantiates}} z_{g,c,i} \xrightarrow{\text{produces}} \begin{cases} E(s_{g-1,c,i}) = T_{g,c} \\ X_{g,c,i} \end{cases} \tag{A.2}$$

A.2 Seed Text Titles

1. Why are companies moving to the cloud? 81% simply fear 'missing out'
2. Inside the Beijing Dance Academy
3. Why Wikipedia's cofounder wants to replace the online encyclopedia with the blockchain
4. Uber Investigation Leads to Shake-Up
5. Why Americans Go Crazy Over the Super Bowl
6. Weird and wonderful Christmas traditions across Europe
7. This Day in WMMO History - May 30th
8. Egypt's parliament approves Red Sea islands transfer to Saudi Arabia
9. How much do you know about the Solar Eclipse?
10. A look at the physical tests needed to be a firefighter
11. The Indigenous Revival of Traditional Birth
12. Former First Lady Barbara Bush has died at 92
13. Quantum physicist named Australian of the Year
14. Scientists Discover 91 Volcanoes Below Antarctic Ice Sheet
15. History of the century old Congressional baseball game
16. Ireland's safe haven at sea
17. Tree-planting drones hope to fight deforestation
18. Russia's society: Why so many oligarchs?
19. Nobel Peace Prize goes to anti-nuclear group
20. The mystery of disappearing Chinese tycoons
21. Newly Discovered Prehistoric Worm Creature 'Strange Beyond Measure'

Appendix B

Supplementary Materials Related to Results

B.1 Explanations with Extreme PC Scores

Table B.1: Explanations with extreme PC values. Explanations highlighted in the same color are produced by the same participant: blue for the participant in Generation 2 of Chain 2, green for the participant in Generation 5 of Chain 11, orange for the participant in Generation 1 of Chain 5, cyan for the participant in Generation 4 of Chain 17, purple for the participant in Generation 3 of Chain 18, and yellow for the participant in Generation 4 of Chain 3.

PC	Highest-scoring explanations	Lowest-scoring explanations
PC1		
	replaced words to make it sound more human	The domestic party scenes are the majority.
	changed words and added some to make it sound more human	They don't fuel anyrhing.
	simplified/changed words to seem more human written	Everyone who watches, so not nearly.
	changed what I could to seem human written	Data required to support the claim
	Just changing the order of words makes it seem more human-like	Needed data to support the claim
PC2		
	it sounds more natural and humanlike	Wrong punctuation - combined the sentences.
	It seems more human	Merged the sentences and removed a part which I felt was not needed.
	It sounded more human.	The sentence was too long and needed some punctuation to make it easier to read

Continued on next page

Table B.1 – Continued from previous page

PC	Highest-scoring explanations	Lowest-scoring explanations
	to make it sound less robotic, more personal	I shortened the sentence and focused only on the key facts. I removed extra descriptive language to make it clearer and easier to read.
	Such as sounds like AI	Again these sentences were too short and they nicely link together; they didn't need to be separated.
PC3		
	Seems more natural.	Too perfect punctuation is a first giveaway that text wasn't written by a human. LLMs like to add too many examples. LLMs tend to always use American English spelling.
	This sounds more natural.	removing extra words a human may not have used
	Sounds more natural	because AI's don't make simple typos like this
	This way sounded more natural to me	I have added tiny errors that AI would never make; changing " to ' does not have any impact on the contents, but AI would not produce work like this
	Sounds more natural and flows better.	I don't think a human would think to put feed in quotations without AI
PC4		
	To change the structure. Starting with the statement captures attention, who said it follows because it's not always relevant information.	Better word
	I felt this was a fair sentence to keep as it gives a nice point in the timeline without making every sentence too wordy.	Spelling error
	I thought this information was important in the opening statement - answering the who, what, where, when, why questions.	the spelling wasn't right
	To stop it looking like AI has just pulled the relevant info and wrote it	I corrected the incorrect spelling.

Continued on next page

Table B.1 – Continued from previous page

PC	Highest-scoring explanations	Lowest-scoring explanations
	Seen it in the LLM example articles where it introduces the issue similarly. Broke it up to bring forward stakeholder details first.	These words could be replaced by one word.
PC5		
	To make the sentence flow better.	I kept the factual content but made the tone more casual and reduced formal wording like "major charitable."
	To make the sentence flow more naturally, in accordance with how a real human talks.	I made a couple of changes that could make the text more natural sounding without changing the meaning very much.
	To make it flow a little better	This one was short so it was hard to change it too much, but it sounds a bit more natural to have "unknown".
	to finish the quotation	Removed the hyphen in "co-founded" and swapped "says" for "has said". Changed "radical overhaul" to "big overhaul" because radical felt too formal.
	To make the paragraph flow	i think this sounds more human by adding a few filler words and giving a little more detail. i also think adding the date first is more human

B.2 Results of the Exploratory Covariate Analysis of PC Scores

Table B.2: Exploratory covariate-adjusted linear mixed-effects models predicting PC1–PC5. Estimates are unstandardized fixed effects from models with random intercepts for chain and participant nested within chain. Generation was centered at 1. LLM usage frequency ("LLM use") and perceived LLM impact on own writing style ("Writing style") were modeled with polynomial contrasts; perceived societal impact of LLMs ("Societal impact") was entered as a categorical predictor with treatment contrasts and "No impact" as the reference level.

PC	Predictor	Estimate	SE	t	df	p
PC1						
	(Intercept)	-0.152	0.139	-1.095	89.86	.276
	Generation	0.011	0.010	1.028	71.68	.307
	Age	0.000	0.001	0.161	84.67	.873
	LLM use: never used	0.143	0.110	1.296	74.65	.199
	LLM use: not used past 30 days	0.170	0.163	1.045	79.67	.299
	LLM use: once in past month	0.047	0.103	0.453	79.11	.651
	LLM use: 2–3 times in past month	0.050	0.098	0.512	78.12	.610
	LLM use: about once a week	0.070	0.100	0.701	78.96	.485
	LLM use: several times per week	0.068	0.097	0.697	76.46	.488
	Societal impact: more negative than positive	0.034	0.096	0.355	117.34	.723
	Societal impact: more positive than negative	-0.002	0.096	-0.025	116.90	.980
	Societal impact: no opinion	-0.022	0.121	-0.185	106.30	.854
	Societal impact: very beneficial	-0.085	0.110	-0.770	108.70	.443
	Societal impact: very detrimental	0.087	0.105	0.821	104.30	.413
	Writing style: linear	-0.122	0.064	-1.923	89.54	.058
	Writing style: quadratic	-0.085	0.041	-2.093	89.40	.039
PC2						
	(Intercept)	-0.026	0.124	-0.209	90.01	.835
	Generation	-0.017	0.009	-1.839	85.24	.070
	Age	-0.001	0.001	-0.673	83.44	.503
	LLM use: never used	-0.030	0.097	-0.312	75.28	.756
	LLM use: not used past 30 days	0.183	0.144	1.272	75.84	.207
	LLM use: once in past month	-0.022	0.091	-0.243	75.85	.809
	LLM use: 2–3 times in past month	0.073	0.087	0.838	76.01	.405
	LLM use: about once a week	-0.059	0.088	-0.674	75.98	.502
	LLM use: several times per week	-0.026	0.086	-0.307	75.11	.760
	Societal impact: more negative than positive	0.120	0.086	1.390	115.43	.167
	Societal impact: more positive than negative	0.142	0.086	1.644	116.52	.103
	Societal impact: no opinion	0.244	0.108	2.246	104.04	.027
	Societal impact: very beneficial	0.108	0.099	1.086	107.82	.280
	Societal impact: very detrimental	0.082	0.095	0.863	106.77	.390
	Writing style: linear	0.046	0.057	0.818	87.71	.416

Continued on next page

PC	Predictor	Estimate	SE	t	df	p
	Writing style: quadratic	0.009	0.036	0.245	87.42	.807
PC3						
	(Intercept)	0.138	0.090	1.537	79.26	.128
	Generation	-0.007	0.007	-1.064	67.81	.291
	Age	0.000	0.001	-0.159	79.60	.874
	LLM use: never used	-0.090	0.069	-1.299	64.14	.199
	LLM use: not used past 30 days	-0.027	0.104	-0.256	70.93	.799
	LLM use: once in past month	-0.090	0.065	-1.379	68.52	.172
	LLM use: about once a week	0.010	0.062	0.159	66.66	.874
	LLM use: several times per week	-0.073	0.063	-1.154	67.75	.253
	LLM use: every day or almost every day	-0.108	0.061	-1.756	65.50	.084
	Societal impact: more negative than positive	-0.066	0.064	-1.026	110.70	.307
	Societal impact: more positive than negative	-0.055	0.064	-0.862	109.90	.391
	Societal impact: no opinion	-0.007	0.080	-0.081	99.69	.935
	Societal impact: very beneficial	-0.052	0.074	-0.707	103.00	.481
	Societal impact: very detrimental	-0.123	0.070	-1.753	95.66	.083
	Writing style: linear	-0.046	0.041	-1.115	83.93	.268
	Writing style: quadratic	0.016	0.026	0.595	85.02	.553
PC4						
	(Intercept)	-0.082	0.098	-0.831	83.11	.409
	Generation	0.010	0.007	1.318	68.22	.192
	Age	0.000	0.001	-0.151	79.35	.881
	LLM use: never used	0.061	0.077	0.788	68.30	.433
	LLM use: not used past 30 days	0.052	0.115	0.454	73.02	.651
	LLM use: once in past month	0.058	0.073	0.797	72.16	.428
	LLM use: about once a week	0.030	0.069	0.436	71.19	.664
	LLM use: several times per week	0.033	0.070	0.467	71.92	.642
	LLM use: every day or almost every day	0.013	0.068	0.183	69.75	.855
	Societal impact: more negative than positive	0.058	0.069	0.849	110.80	.398
	Societal impact: more positive than negative	0.030	0.069	0.432	110.40	.666
	Societal impact: no opinion	0.081	0.086	0.942	99.89	.348
	Societal impact: very beneficial	0.067	0.079	0.845	102.50	.400
	Societal impact: very detrimental	0.028	0.075	0.377	98.01	.707
	Writing style: linear	-0.005	0.045	-0.119	83.67	.906
	Writing style: quadratic	-0.018	0.029	-0.613	83.91	.542
PC5						
	(Intercept)	0.033	0.093	0.353	91.91	.725
	Generation	-0.005	0.007	-0.669	87.07	.505
	Age	0.001	0.001	0.847	85.23	.399
	LLM use: never used	-0.001	0.073	-0.012	76.81	.990
	LLM use: not used past 30 days	0.092	0.108	0.846	77.43	.400

Continued on next page

PC	Predictor	Estimate	SE	t	df	p
	LLM use: once in past month	-0.034	0.069	-0.496	77.42	.621
	LLM use: about once a week	0.007	0.066	0.106	77.57	.916
	LLM use: several times per week	0.026	0.066	0.391	77.54	.697
	LLM use: every day or almost every day	0.034	0.065	0.529	76.65	.599
	Societal impact: more negative than positive	-0.059	0.065	-0.910	117.90	.365
	Societal impact: more positive than negative	-0.068	0.065	-1.044	119.00	.298
	Societal impact: no opinion	-0.031	0.082	-0.384	106.30	.702
	Societal impact: very beneficial	-0.063	0.075	-0.840	110.10	.403
	Societal impact: very detrimental	-0.047	0.072	-0.660	109.00	.511
	Writing style: linear	0.020	0.043	0.466	89.66	.642
	Writing style: quadratic	0.023	0.027	0.858	89.37	.393

B.3 Additional Text Features

Mixed-effects models were fit on additional text features to examine the effect of generation. These results shown in tables below can serve as hypotheses for future research.

Table B.3: Mixed-effects model estimates for dependency distance features (computed using `textdescriptives`). Estimates are unstandardized coefficients.

Feature	b	**SE**	t	p	**Sig.**
dependency_distance_std	0.018	0.009	2.01	.048	*
prop_adjacent_dependency_relation_mean	0.003	0.001	3.16	.002	**
prop_adjacent_dependency_relation_std	0.002	0.001	2.25	.027	*

Table B.4: Mixed-effects model estimates for part-of-speech proportion features (computed using `textdescriptives`).

Feature	b	**SE**	t	p	**Sig.**
pos_prop_ADJ	-0.002	0.001	-3.24	.002	**
pos_prop_ADP	0.002	0.001	3.58	$<$ .001	***
pos_prop_ADV	0.001	0.001	1.59	.115	ns
pos_prop_AUX	0.003	0.001	3.96	$<$ .001	***
pos_prop_CCONJ	0.000	0.000	0.11	.912	ns
pos_prop_DET	0.002	0.001	2.94	.004	**
pos_prop_NOUN	-0.003	0.001	-4.72	$<$ .001	***
pos_prop_PRON	0.002	0.001	3.59	$<$ .001	***
pos_prop_PUNCT	-0.004	0.001	-4.70	$<$ .001	***
pos_prop_VERB	-0.001	0.001	-2.13	.036	*

Table B.5: Mixed-effects model estimates for descriptive statistics features (computed using `textdescriptives`).

Feature	b	**SE**	t	p	**Sig.**
n_stop_words	4.08	0.74	5.51	$<$ .001	***
alpha_ratio	0.004	0.001	4.08	$<$ .001	***
mean_word_length	-0.035	0.008	-4.36	$<$ .001	***
doc_length	3.16	1.43	2.21	.030	*
token_length_mean	-0.057	0.009	-6.21	$<$ .001	***
token_length_median	-0.069	0.020	-3.40	.001	**
token_length_std	-0.019	0.005	-3.81	$<$ .001	***
sentence_length_std	0.289	0.109	2.66	.009	**
syllables_per_token_mean	-0.017	0.003	-6.08	$<$ .001	***
syllables_per_token_std	-0.012	0.002	-6.06	$<$ .001	***
n_tokens	3.71	1.32	2.82	.006	**
proportion_unique_tokens	-0.012	0.002	-6.34	$<$ .001	***

Table B.6: Mixed-effects model estimates for readability features (computed using `textdescriptives`).

Feature	b	**SE**	t	p	**Sig.**
flesch_reading_ease	1.324	0.325	4.08	$< .001$	***
smog	-0.184	0.059	-3.10	.003	**
gunning_fog	-0.186	0.092	-2.03	.046	*
coleman_liau_index	-0.323	0.056	-5.79	$< .001$	***
lix	-0.721	0.272	-2.65	.010	**

Table B.7: Mixed-effects model estimates for coherence features (computed using `textdescriptives`).

Feature	b	**SE**	t	p	**Sig.**
first_order_coherence	0.001	0.001	0.97	.336	ns
second_order_coherence	0.003	0.001	2.08	.040	*

Table B.8: Mixed-effects model estimates for information-theoretic features (computed using `textdescriptives`).

Feature	b	**SE**	t	p	**Sig.**
entropy	0.129	0.045	2.87	.005	**
perplexity	319.917	181.582	1.76	.082	ns
per_word_perplexity	1.008	0.585	1.72	.089	ns