methods

index

the einleitung

inspired by the paper *Empirical evidence of Large Language Model's influence on human spoken communication*, (Yakura et al. (2025)) who indeed found (evidence) for GPT influenced human language after the introduction of chatGPT we tried to replicate the pipeline of building an AI vocabulary (gpt preferred lemmata) and compare frequencies of gpt-typical words across pre- and post chatGPT human language corpora. The first draft essai proves their hypothesis that LLM generated language manifests within human natural language.

preliminary

Our findings are still limited to a yet very small corpus of texts after the introduction of the google gemini chat agent to the public in 03/2024. In contrast to Yakura et al. (2025) and out of resources reasons we decided for gemini as basis for our AI generated vocabulary and for another text corpus (german bundestag plenary protocols, DIP (2026)) than youtube/podcast audio for the same reasons. That limits our post-AI corpus to a small timeframe between 03/2024 up to now. With expanding that corpus to a wider spectrum with including other sources we may harden our results.

hypothesis

following Yakura et al. (2025) we assumed that the consuming of LLM generated language influences the human production of language such that vocabulary typical for LLM output will be found with higher frequencies in human language corpora dating after chat agents introduction.

methods

snc

16062.1

data

our human langugae data consists of raw texts from german bundestag plenary protocols (DIP (2026)). the LLM corpus consists of model summaries of a first subset of these texts generated with the following prompt: Section .

corpus subsets

target	tokens
gemini	3957
human-pre	1514663
human-post	1458323

gemini prompt

```
[1] "System prompt: "
[2] "You are a member of german parliament. Prepare a summary of the text provided to present
[3] "Text:"
```

computation

we first devised AI-typical lemma in the model corpus which are distinctive for that corpus using a linear model (R, lme4::glmer()) that calculates a score for each lemma in the corpus

DIP. 2026. “DIP - Bundestagsprotokolle.” Docs. *DIP - API*. Berlin. https://dip.bundestag.de/%C3%BCber-dip/hilfe/api#content.

Yakura, Hiromu, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. “Empirical Evidence of Large Language Model’s Influence on Human Spoken Communication.” arXiv. https://doi.org/10.48550/arXiv.2409.01754.