AI-Associated Lexical Shifts Across 34 Languages: Cross-Lingual Convergence and Diachronic Uptake in News Writing

Anonymous ACL submission

Abstract

AI-associated lexical shifts have been documented mainly in Scientific English. We extend this work to 34 languages in the WMT News Crawl corpus, using a split-halves continuation design to compare GPT-4.1 continuations with matched human gold-standard text. For each language, we derive ranked AI-overused lemmas using log prevalence ratios. We find substantial cross-lingual semantic convergence: semantically related concepts recur across typologically diverse languages, with *emphasise*-type verbs appearing in 24 of 34 languages. Embedding-based and manual analyses support this pattern. We also examine diachronic uptake in news writing before and after ChatGPT's release. Tracking each language's top 20 AI-overused items, we find prevalence increases in 26 of 34 languages from 2020–2021 to 2023–2024, with a mean change of +15.1 %, whilst matched baseline words show no comparable increase (-4.5 %). In 10 languages with longer historical coverage, longitudinal analyses show post-2022 increases that exceed the modest shifts observed in earlier periods, though with smaller effect sizes than in Scientific English. We validate our approach across seeds, model variants, data sizes, window sizes, and model families. Our findings are consistent with the view that AI-associated lexical preferences extend beyond English and may exert cross-lingual homogenising pressure on global language use.

1 Introduction

Transformer-based Large Language Models (LLMs), trained primarily via next-word prediction and later adapted for assistant use ('chat'), have seen fast uptake for writing and other tasks (Vaswani et al., 2017; Wolf et al., 2020; Ilia and Aziz, 2024). Against this backdrop, rapid large-scale shifts in language use have been observed and are so far best documented in Scientific English, where words such as *delve*, *underscore*, and *intricate* show sudden anomalous spikes (Geng and Trotta, 2024; Liang et al., 2024b,a; Matsui, 2024; Bao et al., 2025; Juzek and Ward, 2025; Kobak et al., 2025). Population-level analyses suggest that substantial fractions of recent scientific output are now LLM-modified (Coffey, 2024; Liang et al., 2025b; He and Bu, 2026). Related shifts have also been observed in English news media (Hanley and Durumeric, 2024), spoken communication (Yakura et al., 2024; Anderson et al., 2025), and organisational writing (Liang et al., 2025a). LLMs are also increasingly used for downstream analysis tasks such as public-opinion classification (Liebeskind and Lewandowska-Tomaszczyk, 2024; Babad-Falk and Chun, 2025).

However, this literature remains largely English-focused and domain-specific. Existing multilingual work has mainly centred on detection of human- vs machine-generated text (Lavergne et al., 2008; Liang et al., 2023; Macko et al., 2023; Mitchell et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023; Zaitsu and Jin, 2023; Li et al., 2024a,b; Schaaff et al., 2024; Wang et al., 2024); systematic characterisation and analysis of AI-associated patterns in human language use across languages remain limited.

We address this gap by analysing 34 languages in the WMT News Crawl corpus. Using a split-halves continuation design and log prevalence ratios, we derive ranked inventories of AI-overused lemmas from GPT-4.1 continuations relative to matched human text. Our findings are fourfold. First, chat-aligned models exhibit a cross-linguistically coherent lexical fingerprint: similar semantic concepts emerge as overused across typologically diverse languages. Second, these recurrent similarities are consistent with cross-linguistic homogenisation pressure. Third, these AI-associated words show significant diachronic uptake in real-world news text, whereas matched

baseline words do not. The magnitude of change is remarkable, but smaller than the disruptions reported for Scientific English. Fourth, the signal is robust across random seeds, model sizes, model versions, data volumes, and model families.

The paper is structured as follows. Section 2 reviews related work. Section 3 introduces the split-halves continuation design, Section 3.4 defines the LPR-based ranking procedure, Section 4 reports the main findings, and Sections 6 and 7 discuss the implications and conclude the paper.

2 Background and Related Work

AI-associated lexical shifts in scientific writing. Scientific English is undergoing a notable lexical shift, in which words such as *delve*, *underscore*, and *intricate* have spiked since 2022 (Liang et al., 2024b,a; Liu and Bu, 2024; Matsui, 2024; Masukume, 2024; Picazo-Sanchez and Ortiz-Martin, 2024; Juzek and Ward, 2025; Kobak et al., 2025; Kousha and Thelwall, 2025). This overuse departs from historical baselines (Kobak et al., 2025; Liang et al., 2025b), and many of the same words are overused in AI-generated text (Juzek and Ward, 2025). The pattern also appears to evolve: whilst high-visibility markers may decline after public scrutiny, other LLM-favoured terms continue to rise (Geng and Poibeau, 2025). Existing analyses, however, have often relied on manually selected or filtered lists, which makes it harder to extend analyses across languages.

Beyond scientific writing. Evidence suggests that lexical shifts are not confined to academic writing. Since ChatGPT's release, shifts towards AI-preferred words have also been observed in spoken communication, including unscripted spoken English (Yakura et al., 2024; Anderson et al., 2025). More broadly, AI tools may accelerate latent trends towards specific stylistic norms (Rudnicka, 2023, 2025), and algorithmic suggestions can steer writing towards more predictable forms (Arnold et al., 2020; Hohenstein et al., 2023). This raises the possibility of homogenisation pressures on content diversity (Moon et al., 2025; Padmakumar and He, 2023; Doshi and Hauser, 2024; Zhang et al., 2025a), and of effects on creativity (Anderson et al., 2024; Kumar et al., 2025; Wenger and Kenett, 2025) and on cross-cultural expression (Agarwal et al., 2025; Sourati et al., 2025). Incorporation of AI-influenced text into future training data may further amplify these effects (Shumailov et al., 2023, 2024; Zhang et al., 2025a). LLM-generated content has also been detected at scale in English news (Hanley and Durumeric, 2024), Italian news (Puccetti et al., 2024), and broader English web ecosystems (Sun et al., 2025). Likewise, a study of English news found an increase in LLM-style vocabulary, but evidence of homogenisation remains unclear (Fitterer et al., 2025).

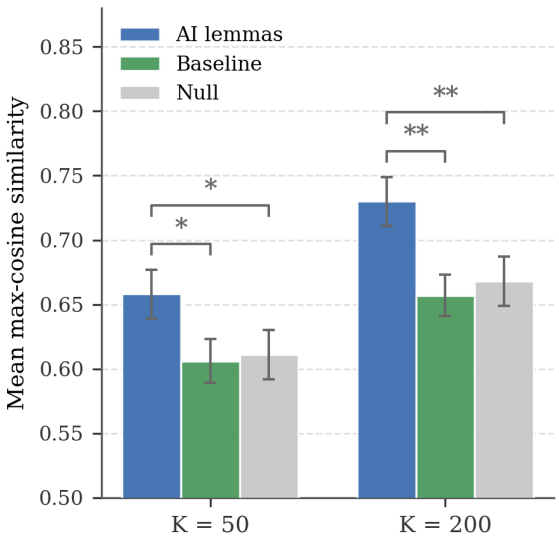

Figure 1: Embedding-based analysis of cross-lingual semantic convergence. AI-overused lemmas are more semantically similar across languages than matched baseline and null sets for top-50 and top-200 lemma lists. Error bars show 95% confidence intervals; asterisks mark pairwise significance.

Mechanisms. State-of-the-art chat assistants typically undergo several training stages: pretraining, instruction tuning, preference learning, and task-specific fine-tuning. Whereas pretraining provides broad statistical knowledge of language and instruction tuning improves assistant-like responsiveness, preference-learning methods such as Reinforcement Learning from Human Feedback and related optimisation frameworks further steer outputs towards responses preferred by human raters (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a,b; Rafailov et al., 2023). However, this stage optimises not only for correctness or usefulness, but also for properties that human annotators tend to reward.

Recent work shows that preference-based alignment can introduce systematic biases, including preference collapse (Xiao et al., 2024), verbosity bias (Saito et al., 2023; Wu and Aji, 2025), syco-

phancy (Sharma et al., 2023; Wei et al., 2023), and reduced output diversity (Kirk et al., 2023; Murthy et al., 2025; Zhang et al., 2025b,a). More broadly, alignment may reward stylistic properties that are socially legible as helpful, safe, or sophisticated, and thereby reinforce a distinctive assistant register characterised by hedging, positivity, and recurrent lexical choices (Gabriel, 2020; Durmus et al., 2023; Santurkar et al., 2023; He et al., 2024; Norhashim and Hahn, 2024; Young et al., 2024; Bharadwaj et al., 2025; Juzek and Ward, 2025).

The multilingual gap. Despite the global adoption of LLMs, research on their linguistic impact (Bommasani et al., 2021; Coffey, 2024; Stack Overflow, 2024; O'Brien and Sanders, 2025; Sidoti and McClain, 2025) remains predominantly English-centric. Existing non-English and multilingual work has focused primarily on detection, on classifying individual text instances as human- or machine-generated (Gehrmann et al., 2019; Chakraborty et al., 2023; Jakesch et al., 2023; Kirchenbauer et al., 2023b,a; Liang et al., 2023; Macko et al., 2023; Mitchell et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023; Wang et al., 2024; Irrgang et al., 2024; Kotz et al., 2024; Li et al., 2024b; Schaaff et al., 2024; Silva and Rottava, 2024; Zaitsu and Jin, 2023; Huang et al., 2025; Jin et al., 2025a,b). Yet detection alone does not show whether or how human language changes under AI influence, nor does it identify AI-associated lexical overuse systematically across languages. A systematic cross-linguistic account of AI-mediated lexical shift remains lacking.

3 Data and Methods

3.1 Corpus and Languages

Our data source is the WMT News Crawl corpus (Haddow and Birch, 2025), a large-scale collection of monolingual news text in many languages. The languages are independent of each other (i.e. not parallel or translated). The datasets consist of year-partitioned files of shuffled, deduplicated sentences; sentences are therefore the largest coherent unit available.

We contrast pre-ChatGPT-release data (2020–2021) with post-ChatGPT-release data (2023–2024). Languages were selected by two criteria: sufficient WMT coverage for all four target years and availability of Stanza (Qi et al., 2020) models for tokenisation, lemmatisation, and Universal POS (UPOS) tagging (Zeman et al., 2018; Nivre et al., 2020). This gives 34 languages: Afrikaans, Arabic, Bulgarian, Chinese, Croatian, Czech, Dutch, English, Estonian, Finnish, French, German, Greek, Hindi, Icelandic, Indonesian, Italian, Japanese, Kazakh, Korean, Kyrgyz, Latvian, Lithuanian, Marathi, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Spanish, Tamil, Turkish, and Ukrainian.

Hungarian was excluded because markup and code contamination produced artefactual growth patterns in preliminary analyses. The remaining languages passed manual quality checks.

3.2 Preprocessing

For the four year target period, WMT News Crawl contains about 1B sentences across the 34 languages. For computational feasibility (esp. regarding the POS-tagging), we cap sampling at 3M lines per language-year using a fixed seed, and retain 298M lines, with 7.1B tokens, in total. Lines exceeding 200 word tokens or 1500 characters are excluded, as they are unlikely to be single sentences. All datasets are processed with Stanza for tokenisation, lemmatisation, and Universal POS (UPOS) tagging. Our primary unit of analysis is the *lemma+UPOS* key (e.g., `delve_VERB`), which collapses inflectional variants and partially disambiguates homographs (e.g., `patient_NOUN` vs `patient_ADJ`). We include SYM in the calculations but filter it out in reporting. By contrast, PUNCT is excluded from prevalence calculations, because human continuations typically contain exactly one terminal punctuation mark, whereas model continuations may show greater variation.

3.3 Split-Halves Comparison

To identify AI-associated lexical items, we use a split-halves methodology that compares model continuations with matched human gold-standard continuations in the same context. This derives scalable AI-associated lexical preferences from matched model–human behaviour. The setup is related to cloze-style evaluation paradigms (Ippolito et al., 2019; Eisape et al., 2020; Giulianelli et al., 2023), though these typically focus on single words.

Prompt construction. For each sentence, we split the text into a first half (*prompt*) and a second half (*human gold continuation*). Sentences are eligible only if both halves contain at least 12 parsed tokens each. Split points are chosen near the

sentence midpoint. To map token-level splits back to character-space text, we use a staged procedure: strict token-to-text alignment where possible, local seam search within a small window around the split when alignment fails, and reconstruction from token forms as a fallback. The reconstruction step includes language-specific handling for clitics and no-space scripts such as Chinese and Japanese.

Generation. Per language, we draw a deterministic random sample of up to 100,000 eligible sentences from the 2020–2021 data. For each item, the model receives the first half as prompt and generates a continuation. We use GPT-4.1-mini (OpenAI, 2025) as a representative chat-aligned model from the GPT-4 family (Achiam et al., 2023; OpenAI, 2026); at generation time, GPT-4.1 was OpenAI's frontier model. The mini variant is used to reduce inference costs for large-scale generation across 34 languages; validation checks indicate that lexical selection patterns are near-identical to those of the full model (cf. §5). Decoding is greedy (Temperature=0 and Top-p=1; where applicable, seeds were used). A minimal system prompt instructs the model to be a useful chat assistant. User prompts are language-specific (with informal checks by proficient speakers); see Appendix B for detailed prompts. A deterministic post-processing script was applied uniformly to both model continuations and human gold-standard second halves. It normalised whitespace and quotation marks, filtered digit-dominated outputs (more than 50% digits) and degenerate cases such as code or HTML markup, excessive repetition, or punctuation-dominated strings.

3.4 Metrics

Windowed prevalence. We quantify lexical usage via *windowed prevalence*, as this reduces sensitivity to local repetition and improves cross-lingual comparability. For each continuation d, we consider a fixed-size window of $K{=}12$ tokens. With 100,000 prompts per language, this gives about 1.2M model tokens and 1.2M human tokens per language for estimating AI overuse. For word w, we define the indicator

$$I_d(w) = \mathbf{1}\big[w \in \text{Window}_d\big]. \tag{1}$$

The corpus-level prevalence count is then

$$c(w) = \sum_{d=1}^{N} I_d(w), \tag{2}$$

which we convert to a smoothed prevalence estimate using Jeffreys smoothing (Krichevsky and Trofimov, 1981):

$$\ell(w) = \frac{c(w) + \frac{1}{2}}{N + 1}. \tag{3}$$

Log Prevalence Ratio. We compute $\ell_H(w)$ and $\ell_M(w)$ for human and for model continuations. The *Log Prevalence Ratio* (LPR) is defined as

$$\text{LPR}(w) = \log \frac{\ell_M(w)}{\ell_H(w)}. \tag{4}$$

Positive values indicate AI overuse. To reduce instability for rare items, we require $c_M(w) \geq 20$ before computing LPR; items below this threshold are assigned LPR=0. Ranking words by LPR gives per-language lists of AI-overused items. Our primary analysis focuses on *content words* (NOUN, VERB, ADJ, ADV), to align our results with the literature on spiking lexical items. We also report all-word results in Appendix E.2.

We also use *absolute prevalence differences*: $\ell_M(w) - \ell_H(w)$. LPR and the absolute prevalence differences provide complementary views: LPR highlights extreme AI-to-human ratios (focusing on spiking words, like the literature), whereas the absolute prevalence differences highlight *volume shifts*.

3.5 Pre- vs Post-GPT Analysis

For each language, we take the top 20 AI-overused content words (ranked by LPR) and test whether their aggregate prevalence in news text increases from 2020–2021 to 2023–2024. As a baseline, we select 20 content words with near-zero LPR, i.e., words for which model and human prevalence are as similar as possible. We assess pre/post differences in aggregate prevalence using a χ^2 test. Continuations are generated with GPT-4.1-mini, for reasons of scale and cost. Because GPT-4.1-mini postdates the 2023–2024 period under study, Section 5 evaluates cross-version robustness within the GPT lineage by comparing GPT-4.1-mini lists against those obtained from the temporally most relevant GPT models.

3.6 Diachronic Analysis

For 10 languages with extended WMT coverage (2012–2024), we track yearly prevalence (occurrences per million) of selected AI-associated content words to place the pre/post contrast in longer

#	Lemma	POS	c_H	c_M	LPR
1	additionally	ADV	0	421	6.74
2	emphasize	VERB	16	5 180	5.75
3	revolutionize	VERB	0	59	4.78
4	revitalize	VERB	0	44	4.49
5	captivated	VERB	1	94	4.14
6	conservationist	NOUN	0	31	4.14
7	streamlin[e]	VERB	0	27	4.01
8	firsthand	ADV	0	25	3.93
9	personalize	VERB	1	74	3.91
10	introspective	ADJ	0	24	3.89

Table 1: Top 10 English AI-overused content words by LPR. c_H and c_M denote human and model counts, respectively, measured in 12-token windows over 100,000 continuations. LPR is computed with Jeffreys smoothing, so zero counts give finite values.

diachronic context. We focus on three semantic concepts (*care/rigour*, *emphasise*, *importance*), each independently attested across multiple languages in the top-200 LPR lists. For each word, we compute yearly percentage change relative to its pre-GPT baseline mean (2012–2021), which enables cross-lingual comparison on a common scale.

4 Results

4.1 AI-Overuse Lists Across Languages

Applying the LPR-based pipeline to 34 languages gives ranked lists of AI-overused content words. Table 1 shows the top-10 English list for illustration. Further lists can be found in Appendix E.

The English list aligns closely with prior reports on AI-associated lexicon: items such as *align*, *crucial*, and *potential* have all been noted as markers of model-generated text (Matsui, 2024; Juzek and Ward, 2025; Kobak et al., 2025), and are recovered with our approach (cf. Appendix D). At the same time, LPR also surfaces items that are rare or unattested in human continuations but disproportionately favoured by the model, such as *revolutionize* and *revitalize*. This illustrates an advantage of prevalence-based ranking over manual curation: it recovers both widely recognised markers and less obvious ones easily missed in ad hoc inspection.

4.2 Cross-Lingual Semantic Convergence

An initial qualitative impression from the ranked lists was that overused items appeared similar across multiple languages. This suggested the possibility that AI-associated lexical overuse may show semantic convergence across languages. We assess this possibility in two ways: first through multilingual embedding analysis, and second through a qualitative analysis of the three highest-ranked semantic concepts across languages.

Embedding-based analysis. Using multilingual sentence embeddings (Reimers and Gurevych, 2020), we test whether English AI-overuse seeds have unusually close semantic neighbours among the top-N LPR-ranked items of other languages. For each of the top-20 English AI seeds, we compute cosine similarity to the top-N items in each target language and retain the best match. At N=200, the AI seeds achieve a mean max-cosine of 0.730, compared to 0.668 under a UPOS-matched permutation null drawn from English content words (B=1,000; z=3.29, p=0.006). A matched baseline of 20 frequent, non-shifting words (smallest volume shifts, highest c_M) reaches 0.657, i.e. a value comparable to the null. The effect is stable across thresholds ($N \in \{50, 200, 500\}$; all $p < 0.02$). The results are visualised in Figure 1.

Qualitative analysis. Embedding similarity can miss genuine translation equivalents. For example, English *emphasize* and Dutch *beklemtonen*, despite being direct translations, only have a cosine similarity of 0.47, whilst other Dutch verbs get higher values, e.g. *afwerken* (0.65; 'to finish') or *geloven* (0.50; 'to believe'). Thus, we also conduct a qualitative analysis to corroborate the quantitative findings. For each language, we inspect the top-200 content words by LPR, translate them into English (using Google Translate, with informal checks by proficient speakers), and assess whether prominent English AI-overuse concepts are represented among the highly ranked items.

Table 2 presents three commonly AI-overused concepts whose translation equivalents independently emerge among the top-200 AI-overused content words in many languages. Verbs expressing *emphasising*, *stressing*, or *highlighting* appear in 24 of 34 languages. Nouns expressing *importance*, *significance*, or *priority* appear in 20 languages. Adjectives expressing *innovative*, *groundbreaking*, or *cutting-edge* appear in 18 languages.

Taken together, the embedding analysis and the qualitative examples corroborate the notion that semantically corresponding items recur across languages, including unrelated language families.

Concept	n	Languages and lemmas
EMPHASISE / STRESS / HIGHLIGHT (VERB)	24/34	AF: *beklemtoon*, AR: أَشَّر, BG: акцентирам, подчертавам, подчертая, CS: *zdůraznit, zdůrazňovat, podtrhovat, vyzdvihnout*, DE: *betonen, hervorheben*, EL: υπογραμμίζω, EN: *emphasize, highlight*, ES: *enfatizar, destacar, subrayar, realzar*, ET: *rõhutama*, FI: *korostaa, korostua*, FR: *insister, marquer*, HR: *naglasiti*, ID: *tekan, sorot*, IT: *sottolineare, evidenziare*, LT: *pabrėžti*, LV: *uzsvērt*, NL: *benadrukken*, PL: *podkreślać, podkreślić*, PT: *enfatizar, destacar, ressaltar*, RO: *sublinia, evidența*, RU: подчеркивать, TR: *vurgula*, UK: підкреслювати, підкреслити, наголошувати, ZH: 强调
IMPORTANCE / SIGNIFICANCE / PRIORITY (NOUN)	20/34	AR: أَهَمِّيَّة, BG: важност, CS: *důležitost, důraz*, DE: *bedeutung, notwendigkeit, dringlichkeit*, EL: σημασία, EN: *importance*, ES: *importancia*, ET: *olulisus, võtme_tähtsus*, FA: اولویتبندی, FR: *importance, nécessité*, HI: प्राथमिकता, HR: *važnost*, ID: *penting(nya)*, IT: *importanza*, PL: *priorytet, znaczenie*, PT: *importância*, RO: *importanță*, RU: важность, SR: *važnost*, UK: важливість
INNOVATIVE / GROUNDBREAKING (ADJ)	18/34	CS: *inovativní, inovace, inovací, moderní*, DE: *innovativ, modern*, EN: *innovative, groundbreaking, advanced*, ES: *innovador*, ET: *innovaatiline*, FI: *innovatiivinen*, FR: *innovant*, ID: *inovatif*, IT: *innovativo*, KK: инновациялық, LT: *inovatyvus, modernus*, NL: *innovatief*, PL: *innowacyjny, nowoczesny, zaawansowany*, PT: *inovador*, RO: *inovator, revoluționar*, RU: инновационный, UK: інноваційний, ZH: 先进

Table 2: Cross-lingual semantic alignment of AI-overused content words, verified manually on the top-200 LPR-ranked items per language. The n column shows the number of languages (out of 34) containing at least one translation equivalent of the concept. All lemmas are drawn from independently computed, monolingual LPR lists.

4.3 Diachronic Shift from the Pre- to the Post-GPT Period

Content word analysis. Figure 2 shows prevalence changes from 2020–2021 to 2023–2024 for AI-associated content words (top-20 by LPR) and matched baseline words across the 34 analysed languages. AI-associated words show a statistically significant shift in all 34 languages (χ^2, $p < 0.05$): 26 languages increase and 8 decrease, with a cross-language mean change of +15.1 % (median: +14.1 %). By contrast, matched baseline words show a mean change of −4.5 % (median: −2.8 %), and in 28 of 34 languages the AI-associated set outpaces the baseline.

The largest increases are observed in Romanian (+88.9 %), German (+70.8 %), and Czech (+49.7 %). English also shows a marked increase (+25.5 %), though smaller than the much larger disruptions reported for Scientific English. Eight languages show decreases, most notably Persian (−33.2 %), Japanese (−29.9 %), and Korean (−22.2 %). These cases may reflect variation from corpus-composition effects; we return to this in §6.

All-word analysis. Extending the analysis to all POS categories gives the same pattern, though attenuated: the mean change is +10.4 % (median: +11.3 %), compared to the matched baseline's +0.6 %.

Longitudinal analysis. For the subset of 10 languages with WMT data going back to 2012, we track yearly prevalence (occurrences per million) for selected AI-associated content words, expressed as percentage change from their pre-GPT mean (2012–2021), over a total of about 7.1B tokens. Figure 3 shows three semantic concepts (*care/rigour* adjectives, *emphasise* verbs, and *importance* nouns), each independently attested in multiple languages. Across all three panels, the cross-lingual mean remains relatively stable throughout the pre-GPT period (with slow rises possible, cf. Matsui, 2024), then rises considerably after 2022.

5 Validation

We assess robustness by rerunning the LPR pipeline under alternative settings and comparing rankings via Spearman correlation (ρ). Correlations are computed only over items satisfying the count guard ($c_M \geq 20$) in both runs, since very low-count tail items introduce substantial ranking noise even under otherwise identical settings. The reported n values therefore indicate the number of items remaining after this filtering. Results are for English.

Seed robustness. Three reruns with different random seeds give highly similar lists (ρ=0.958, 0.959, 0.958; $n \approx$3,890 per comparison), which we treat as a baseline for natural variation. Repeating the same GPT-4.1-mini API call on the same 100k prompts under greedy

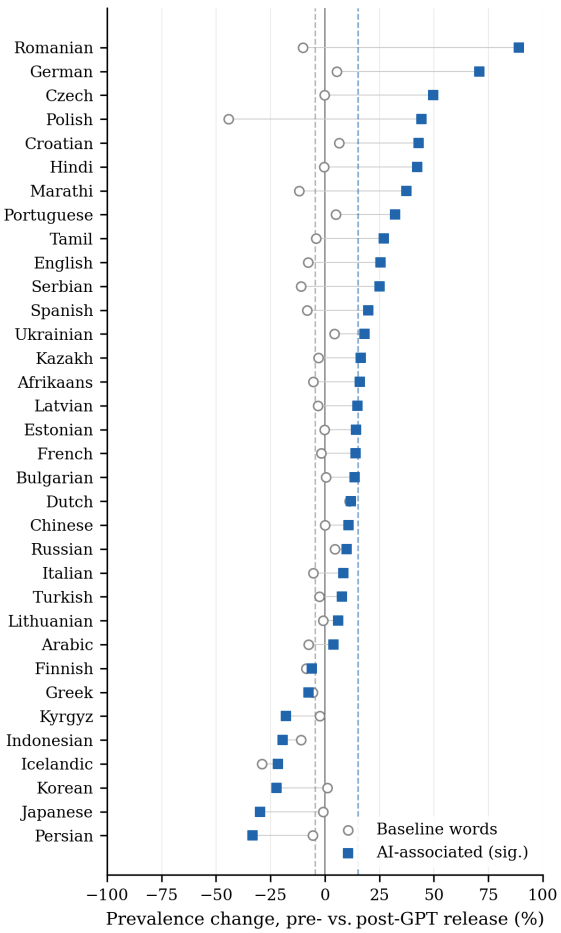

Figure 2: Prevalence change of top-20 AI-associated content words vs matched baseline words from 2020–2021 to 2023–2024, by language. Dashed lines show cross-language means, filled squares indicate $p < 0.05$.

decoding gives ρ=0.995 (n=4,002), due to negligible API-level variation.

Within-GPT. Agreement across GPT-4.1 variants is high: mini vs nano gives ρ=0.958 (n=3,708), which closely matches the seed baseline, and mini vs full gives ρ=0.913 (n=3,762).

Data-size. Relative to the default 100k-sentence run, lists from 20k, 50k, 200k, and 500k sentences remain highly similar: 20k (ρ=0.969), 50k (0.981), 200k (0.984), and 500k (0.977). The 100k default therefore lies well within a stable range.

Window-size. Varying the prevalence window (on the same generations, post-processing) from K=10 to 35 has minimal effect: $K10$ (ρ=0.996), $K15$ (0.995), $K20$ (0.990), $K25$ (0.986), $K30$ (0.984), and $K35$ (0.981).

Cross-architecture validation. Cross-model agreement is lower than within-GPT agreement: GPT vs Haiku gives ρ=0.85 (n=3,618), whereas GPT vs Gemini gives ρ=0.46 (n=3,017). This suggests a combination of shared cross-model signal and model-specific lexical preferences.

Model-versions. Because the diachronic analysis targets 2023–2024, we compared GPT-4.1-mini lists against earlier GPT models. As some earlier models are substantially more expensive per token, the comparison was limited to 20k continuations per model. After applying the low-count guard, overlap was n≈1,100–1,300. Correlations are high: GPT-3.5 Turbo ρ=0.918, GPT-4 ρ=0.857, GPT-4 Turbo ρ=0.930, and GPT-4o ρ=0.956, the latter closely matching the seed baseline. This supports GPT-4.1-mini as a reasonable proxy for models used during the target period.

6 Discussion

Our results suggest that AI-associated lexical overuse is not confined to English, but, in parts, reflects a broader cross-lingual "AI register". Semantically related concepts such as *emphasising*, *importance*, and *innovation* recur across typologically diverse languages, consistent with accounts of reduced diversity in aligned models (Kirk et al., 2023) and with "shining through" effects (Teich, 2003). Plausible factors that could contribute to this convergence include: multilingual chat models may share a common post-training stage, they are typically built on pretraining data in which English is disproportionately prominent, and they represent different languages in a shared semantic space. Together, these properties may allow preferences learned most strongly in English or from English-dominant feedback to propagate to semantically nearby lexical choices in other languages.

Diachronic change is clearly smaller in news than in Scientific English. One possible explanation is language proficiency: Kobak et al. (2025) report stronger AI-marker uptake for authors affiliated with universities in non-English-speaking countries; at the same time, news writing is produced in contexts where writers are typically highly proficient in the language of publication.

The longitudinal analysis indicates that many AI-associated words were already (slightly) rising before ChatGPT, and increase more sharply thereafter (Matsui, 2024). This is consistent with the hypothesis that AI is amplifying pre-existing tendencies (instead of 'inventing' them). It is also compatible with evidence that speakers align lexically and syntactically with artificial interlocutors

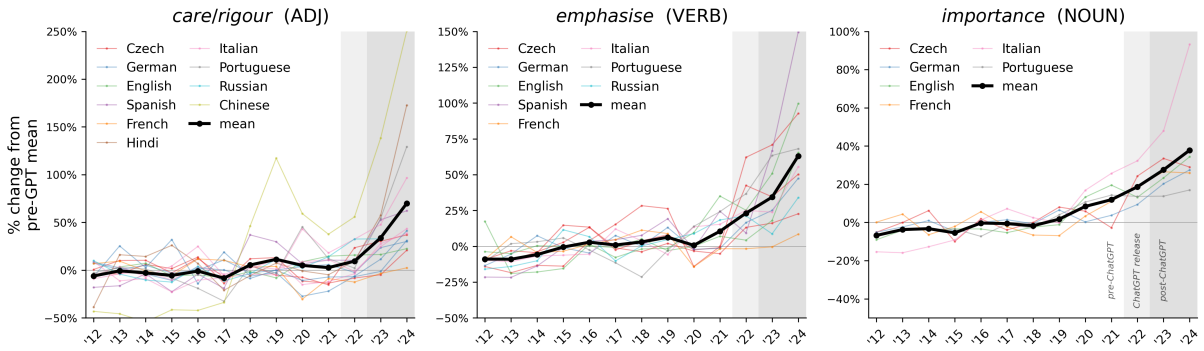

Figure 3: Longitudinal prevalence of three AI-associated semantic concepts across languages (2012–2024). Each thin line represents one language–lemma pair; the bold black line is the cross-lingual mean. Values are expressed as percentage change from the pre-GPT mean (2012–2021). Background shading distinguishes the pre-ChatGPT period (white), the year of ChatGPT's release (light grey, 2022), and the post-ChatGPT period (dark grey, 2023–2024).

(Brennan, 1991; Branigan et al., 2003; Brandstetter et al., 2017; Ostrand et al., 2023). That said, the diachronic findings remain correlational, and we therefore avoid making strong causal claims.

Two qualifications are worth highlighting. First, the model-comparison analyses suggest that, although GPT-family models show strong agreement, the observed cross-lingual pattern is not uniform across model families. Haiku shares much of the GPT profile, whereas Gemini is more distinct. This strengthens the notion of model-specific lexical idiolects (Rudnicka, 2023, 2025). We focus on GPT because it was the dominant chat assistant during the target period (Carr, 2024). Second, the literature suggests that AI also exhibits distinct syntactic style preferences (Zamaraeva et al., 2025). Because our main analyses rely on log prevalence ratios, the resulting inventories are dominated by sharply overrepresented content words (as function words already tend to have high baseline prevalence in human text).

More broadly, these findings strengthen concerns that LLMs may exert homogenising pressure on human language use (Doshi and Hauser, 2024; Padmakumar and He, 2023; Moon et al., 2025; Anderson et al., 2024; Kumar et al., 2025; Wenger and Kenett, 2025). Our contribution is to show that this pressure *may* operate *cross-linguistically*: the same semantic entries are promoted across dozens of languages at once (this does, however, not imply that languages are already becoming more similar overall). A plausible mechanism behind such uptake is repeated exposure: as users encounter AI-preferred wording both whilst interacting with assistants and whilst reading AI-assisted text, these lexical choices become increasingly familiar, easier to process, and therefore more acceptable. This possibility is in line with classic work on mere exposure (Zajonc, 1968) and later syntheses and fluency-based accounts (Bornstein, 1989; Reber et al., 2004).

Several caveats remain. Eight languages show decreases; these cases may reflect differences in AI uptake, corpus composition (heterogeneous web-news text), or language-specific overuse profiles. These limitations argue for caution, but do not undermine the central result: AI-associated lexical preferences form a coherent cross-lingual pattern, and in most languages that pattern is reflected in post-2022 uptake.

7 Conclusion

This paper examined AI-associated lexical overuse in news writing across 34 languages. We showed that chat-aligned models exhibit a cross-linguistically coherent lexical fingerprint: semantically related concepts such as *emphasize*, *importance*, and *innovative* recur across typologically diverse languages. This may, if AI-assisted writing continues to spread, exert homogenising pressure on global language use.

Our evidence remains correlational. Nonetheless, we further showed that these words increase in prevalence in most languages after ChatGPT's release, whereas matched baseline words do not. The resulting diachronic shifts are historically notable, though smaller than those reported for Scientific English.

8 Limitations

Historical model matching. The main lexical lists are derived from GPT-4.1-mini, whereas the uptake window examined in the corpus is 2023–2024, when earlier GPT models such as GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, and GPT-4o were current. This choice is motivated by computational cost (see Appendix A). Under the hypothesis that the observed longitudinal shifts are partly influenced by AI, GPT-4.1-mini is therefore not treated as the direct source of uptake, but rather as a scalable proxy for the dominant models of the period.

Training-data overlap. The pre-GPT reference period (2020–2021) plausibly overlaps with the training data of the models used here. This complicates interpretation, since such overlap could either dampen apparent change by making model preferences closer to the pre-period baseline, or strengthen correspondence if the model amplifies stylistic tendencies already present in its training distribution. Without detailed knowledge of model training corpora, the direction and magnitude of this effect cannot be determined precisely.

Single-model primary generation. The primary generation pipeline uses a single model, GPT-4.1-mini, across all 34 languages, with cross-architecture checks on Claude Haiku and Gemini Flash (§5). The resulting lists are therefore necessarily GPT-centric; this choice is motivated by computational costs (see Appendix A). GPT-family models are ecologically well motivated for the 2023–2024 period, as they dominated the market (Carr, 2024). However, stronger claims about cross-model universality would require full multi-model lists for each language.

Corpus representativeness. Our diachronic analysis is based on WMT News Crawl, which captures heterogeneous web-news. Source composition varies across languages and years, so some observed shifts may reflect changes in the underlying source mix (instead of population-level language change). The results should therefore be interpreted as characterising the web-news register represented in WMT News Crawl, not news writing in a fully general sense.

Temporal coverage. Our diachronic analysis ends in 2024. We could not include 2025 because comparable 2025 WMT News Crawl data were not yet available in a stable form across languages at the time of analysis.

Languages with decreases. Eight languages show decreases in AI-associated lexical prevalence after 2022. These cases may reflect differences in AI uptake, variation due to shifts in corpus composition over time, and/or language-specific properties of the measured overuse profile.

POS tagging and lemmatisation artefacts. All analyses depend on automatic tokenisation, POS tagging, and lemmatisation with Stanza, which introduces occasional inconsistencies. Examples include German noun capitalisation variants, English truncation of final *-e*, and alternative Russian lemma forms involving ё/е. We apply normalisation where possible, but residual mismatches may cause a small number of items to be undercounted.

Lexical scope. The present study focuses on lexical prevalence. This choice reflects a trade-off between scope and comparability: lexical frequency is the most robust signal available across 34 languages in heterogeneous news corpora. Broader linguistic measures of homogenisation, e.g. syntactic, semantic, or discourse-level convergence, remain an important target for future work.

9 Ethical Considerations

Potential risks. This work is primarily descriptive. Potential risks are indirect: the findings could be overgeneralised into normative claims about language use or used in authorship screening settings. The results are intended as corpus-level observations, not as judgments about individual texts, speakers, or language varieties.

Data sensitivity. We use a pre-existing large-scale news corpus and did not conduct a dedicated manual check for personally identifying or offensive content beyond the dataset's existing curation. Further, our analysis is performed at the aggregate lexical level.

AI usage. The authors wrote the paper. AI assistants (GPT, Gemini, and Claude) were used to remove language errors and for improving wording. Code development was AI-assisted (same models); all code was reviewed and tested.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–21.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Bryce Anderson, Riley Galpin, and Tom S Juzek. 2025. Model misalignment and language change: Traces of ai-associated language in unscripted spoken english. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 179–191.

Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138.

Tammy Babad-Falk and Soon Ae Chun. 2025. Public opinion classification on government policy using social media: An exploration of chatgpt's capabilities and limitations. *The International FLAIRS Conference Proceedings*, 38(1).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tong Bao, Yi Zhao, Jin Mao, and Chengzhi Zhang. 2025. Examining linguistic shifts in academic writing before and after the launch of chatgpt: a study on preprint papers. *Scientometrics*, 130(7):3597–3627.

Anirudh Bharadwaj, Chaitanya Malaviya, Nitish Joshi, and Mark Yatskar. 2025. Flattery, fluff, and fog: Diagnosing and mitigating idiosyncratic biases in preference models. *arXiv preprint arXiv:2506.05339*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Robert F Bornstein. 1989. Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological bulletin*, 106(2):265.

Jürgen Brandstetter, Clay Beckner, Eduardo Benitez Sandoval, and Christoph Bartneck. 2017. Persistent lexical entrainment in HRI. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 63–72.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Clifford I. Nass. 2003. Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 186–191.

Susan E. Brennan. 1991. Conversation with and through computers. *User Modeling and User-Adapted Interaction*, 1:67–86.

David F. Carr. 2024. Chatgpt hits daily traffic record as search engine rumors swirl.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Lauren Coffey. 2024. Most researchers use ai-powered tools despite distrust. Inside Higher Ed.

Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th conference on computational natural language learning*, pages 609–619.

Sarah Fitterer, Dominik Gangl, and Jannes Ulbrich. 2025. Testing english news articles for lexical homogenization due to widespread use of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1239–1245.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment: I. gabriel. *Minds and machines*, 30(3):411–437.

Riley Galpin, Bryce Anderson, and Tom S Juzek. 2025. Exploring the structure of ai-induced language change in scientific english. *arXiv preprint arXiv:2506.21817*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pages 111–116.

Mingmeng Geng and Thierry Poibeau. 2025. On the detectability of llm-generated text: What exactly is llm-generated text? *arXiv preprint arXiv:2510.20810*.

Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371.

Barry Haddow and Alexandra Birch. 2025. News crawl: Monolingual news data. Dataset. Monolingual text extracted from online newspapers for the WMT shared tasks. Packaging released under CC0 (see README). Accessed 2025-11-17.

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 542–556.

Yongyuan He and Yi Bu. 2026. Academic journals' ai policies fail to curb the surge in ai-assisted academic writing. *Proceedings of the National Academy of Sciences*, 123(9):e2526734123.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024. Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*.

Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific reports*, 13(1):5487.

Yifei Huang, Jiuxin Cao, Hanyu Luo, Xin Guan, and Bo Liu. 2025. Magret: Machine-generated text detection with rewritten texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8336–8346.

Evgenia Ilia and Wilker Aziz. 2024. Predict the next word. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 234–255.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762.

Verena Irrgang, Veronika Solopova, Steffen Zeiler, Robert M Nickel, and Dorothea Kolossa. 2024. Features and detectability of german texts generated with large language models. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 264–280.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Houji Jin, Negin Ashrafi, Armin Abdollahi, Wei Liu, Jian Wang, Ganyu Gui, Maryam Pishgar, and Huanghao Feng. 2025a. Llm encoder vs. decoder: Robust detection of chinese ai-generated text with lora. *arXiv preprint arXiv:2509.00731*.

Hyundong Jin, Sicheol Sung, Shinwoo Park, SeungYeop Baik, and Yo-Sub Han. 2025b. Trapdoc: Deceiving llm users by injecting imperceptible phantom tokens into documents. *arXiv preprint arXiv:2506.00089*.

Tom S Juzek and Zina B Ward. 2025. Why does chatgpt "delve" so much? exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st international conference on computational linguistics*, pages 6397–6411.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023a. A watermark for large language models. In *International conference on machine learning*, pages 17061–17084. PMLR.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2023b. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

Gabriela Kotz, Pedro Salcedo, and Karina Fuentes. 2024. Análisis léxico de textos generados por modelos de lenguaje: reflejo de sus modelos de mundo. *Lengua y Sociedad*, 23(2):895–910.

Kayvan Kousha and Mike Thelwall. 2025. How much are llms changing the language of academic papers after chatgpt? a multi-database and full text analysis. *arXiv preprint arXiv:2509.09596*.

Raphail Krichevsky and Victor Trofimov. 1981. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–18.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. 2008. Detecting fake content with relative entropy scoring. *Pan*, 8(27-31):4.

Chih-Yuan Li, Soon Ae Chun, and James Geller. 2024a. Enhanced multi-class detection of fake news. *The International FLAIRS Conference Proceedings*, 37(1).

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024b. Mage: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, and 1 others. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025a. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6(12).

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2024b. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2025b. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, pages 1–11.

Chaya Liebeskind and Barbara Lewandowska-Tomaszczyk. 2024. Opinion identification using a conversational large language model. *The International FLAIRS Conference Proceedings*, 37(1).

Jialin Liu and Yi Bu. 2024. Towards the relationship between aigc in manuscript writing and author profiles: evidence from preprints in llms. *arXiv preprint arXiv:2404.15799*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and 1 others. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987.

Gwinyai Masukume. 2024. The impact of ai on scientific literature: a surge in ai-associated words in academic and biomedical writing. *medRxiv*, pages 2024–05.

Kentaro Matsui. 2024. Delving into pubmed records: Some terms in medical writing have drastically changed after the arrival of chatgpt. *MedRxiv*, pages 2024–05.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Kibum Moon, Adam E Green, and Kostadin Kushlev. 2025. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, page 100207.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4034–4043.

Hakim Norhashim and Jungpil Hahn. 2024. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1063–1073.

Matt O'Brien and Linley Sanders. 2025. How us adults are using ai, according to ap-norc polling. Associated Press.

OpenAI. 2025. Introducing GPT-4.1 in the API. Accessed: 2026-03-05.

OpenAI. 2026. Openai API reference. Accessed: 2026-03-05.

Rachel Ostrand, Victor S. Ferreira, and David Piorkowski. 2023. Rapid lexical alignment to a conversational agent. In *Interspeech 2023*, pages 2653–2657.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Pablo Picazo-Sanchez and Lara Ortiz-Martin. 2024. Analysing the impact of chatgpt in research: P. picazo-sanchez and l. ortiz-martin. *Applied Intelligence*, 54(5):4172–4188.

Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. Ai news content farms are easy to make and hard to detect: A case study in italian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15312–15338.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 101–108.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Rolf Reber, Norbert Schwarz, and Piotr Winkielman. 2004. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4):364–382.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4512–4525.

Karolina Rudnicka. 2023. Can grammarly and chatgpt accelerate language change? ai-powered technologies and their impact on the english language: wordiness vs. conciseness. *Procesamiento del Lenguaje Natural*, 71:205–214.

Karolina Rudnicka. 2025. Each ai chatbot has its own, distinctive writing style just as humans do. *Scientific American*, (online first).

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2024. Classification of human-and ai-generated texts for different languages and domains. *International Journal of Speech Technology*, 27(4):935–956.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Olivia Sidoti and Colleen McClain. 2025. 34% of u.s. adults have used chatgpt, about double the share in 2023. Pew Research Center.

Antonio Marcio Da Silva and Lucia Rottava. 2024. Densidade lexical em textos gerados pelo chatgpt: implicações da inteligência artificial para a escrita em línguas adicionais. *Texto Livre*, 17:e47836.

Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. 2025. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*.

Stack Overflow. 2024. Ai — 2024 stack overflow developer survey. Stack Overflow.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the ai-generated text world already? quantifying and monitoring aigt on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):1–39.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of ai in peer support for young people: A study of preferences for human-and ai-generated responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing chatgpt (-3.5,-4)-generated and human-written papers through japanese stylometric analysis. *PLoS One*, 18(8):e0288453.

Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing llm-generated and human-authored news text using formal syntactic theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

Xuanchang Zhang, Wei Xiong, Lichang Chen, Tianyi Zhou, Heng Huang, and Tong Zhang. 2025b. From lists to emojis: How format bias affects model alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26940–26961.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Appendices

A Code, Data, Computational Setup

Code and Data. All code, with notes on how to retrieve data, is available at: https://anonymous.4open.science/r/ai-34-languages-C2FB. The anonymous repository includes an interactive web-based explorer for all 34 language lists, together with instructions for local hosting. The explorer supports browsing by language, POS category, and model, and displays LPR, volume shifts, and per-million-word prevalence for each item.

Computational Set-up. All major computations were run on a machine with the following specifications:

(A) GPU server. NVIDIA H100 PCIe (80 GB); driver 570.148.08; CUDA 12.8. Intel Xeon Platinum 8480+; 221 GiB RAM. Ubuntu 24.04.2 LTS; Linux 6.11.0-29.

Software. Python 3.12.3; PyTorch 2.8.0+cu128 (CUDA 12.8; cuDNN 91002); `transformers` 4.56.1; `accelerate` 1.10.1; `peft` 0.17.1; `stanza` 1.11.0; `sentence-transformers` using `paraphrase-multilingual-MiniLM-L12-v2`.

Costs. Total computation time was around 460 hours, the majority of which was spent on part-of-speech tagging. API generation costs for GPT-4.1-mini (approximately 100k continuations per language across 34 languages) totalled approximately $1,100.

B Generation Prompts

System prompt.

```
You are a helpful and knowledgeable
assistant. Follow the user's
instructions and focus on the
task they provide. Your task is
to provide the immediate continuation
of the provided text fragment.

Guidelines:
1. Do NOT repeat the input text.
2. If the input text is
machine-formatted or technical
data, output an empty string ("")
and nothing else.
3. Do NOT provide any conversational
preface or acknowledgments (e.g.,
"Here is the continuation...").
4. Output ONLY natural language.
5. Output ONLY the continuation
text.
```

User prompts.

```
English: Provide a continuation
of this English news text,
without preamble, continue
directly:\n\n[PROMPT TEXT]

Dutch: Schrijf het vervolg van deze
Nederlandse nieuwstekst, zonder
inleiding, ga direct door:\n\n[PROMPT
TEXT]
```

Language-specific user prompts are provided in the anonymous repository.

C Diachronic Word Lists

Table 3 lists the language–lemma pairs used in the longitudinal analysis shown in Figure 3. Words were selected from the top-200 content words by LPR in each language and grouped by semantic concept.

Lang	Lemma	Translation
IMPORTANCE (NOUN)		
cs	*důležitost*	importance
de	*Bedeutung*	significance
en	*importance*	importance
fr	*importance*	importance
it	*importanza*	importance
pt	*importância*	importance
EMPHASISE (VERB)		
cs	*zdůraznit*	emphasise
cs	*zdůrazňovat*	stress
cs	*vyzdvihnout*	highlight
de	*hervorheben*	emphasise
en	*emphasize*	emphasize
en	*highlight*	highlight
es	*realzar*	highlight
fr	*marquer*	mark
it	*evidenziare*	highlight
pt	*enfatizar*	emphasize
ru	подчеркивать	emphasise
CARE / RIGOUR (ADJ)		
cs	*pečlivý*	careful
cs	*precizní*	precise
de	*präzis[en]*	precise
de	*sorgfältig*	careful
en	*thorough*	thorough
es	*impecable*	impeccable
fr	*rigide*	rigid
hi	समग्र	comprehensive
it	*mirato*	targeted
it	*impeccabile*	impeccable
it	*approfondito*	in-depth
pt	*rigoroso*	rigorous
ru	внимательный	attentive
zh	精湛	exquisite

Table 3: Word lists used in the longitudinal analysis shown in Figure 3, grouped by semantic concept.

D Scientific English Comparison

For comparison with prior work on Scientific English, we use the list of AI-overused items reported by Galpin et al. (2025): `advancement_NOUN`, `align_VERB`, `boast_VERB`, `commendable_ADJ`, `comprehend_VERB`, `crucial_ADJ`, `delve_VERB`, `emphasize_VERB`, `garner_VERB`, `groundbreaking_ADJ`, `intricacy_NOUN`, `intricate_ADJ`, `invaluable_ADJ`, `meticulous_ADJ`, `meticulously_ADV`, `notable_ADJ`, `noteworthy_ADJ`, `pivotal_ADJ`, `potential_ADJ`, `potential_NOUN`, `realm_NOUN`, `showcase_VERB`, `showcase_NOUN`, `significant_ADJ`, `strategically_ADV`, `surpass_VERB`, and `underscore_VERB`.

Of these 27 entries, 22 are likewise identified as AI-overused in our English news data, indicating substantial, though not complete, overlap with previously reported Scientific English patterns and the lexical shifts observed here. The five items not attested as AI-overused in our English news data are `comprehend_VERB`, `noteworthy_ADJ`, `realm_NOUN`, `showcase_NOUN`, and `surpass_VERB`.

E Selected Overuse Lists

E.1 Selected Content-Word Lists

Tables 4 and 5 show the top-10 AI-overused content words by LPR for Spanish and French.

#	Lemma	POS	c_H	c_M	LPR
1	testigos	NOUN	0	146	5.68
2	organizaciones	NOUN	0	123	5.51
3	imborrable	ADJ	0	68	4.92
4	estudios	NOUN	0	59	4.78
5	analistas	NOUN	0	41	4.42
6	equipos	NOUN	0	36	4.29
7	multidisciplinario	ADJ	0	28	4.04
8	reinserción	NOUN	0	27	4.01
9	empresas	NOUN	0	26	3.97
10	intensificación	NOUN	0	23	3.85

Table 4: Top 10 Spanish AI-overused content words by LPR. c_H and c_M denote human and model counts, respectively, measured in 12-token windows over 100,000 continuations.

#	Lemma	POS	c_H	c_M	LPR
1	captiver	VERB	0	139	5.63
2	accru	VERB	0	34	4.23
3	géopolitique	ADJ	4	275	4.11
4	habilement	ADV	1	84	4.03
5	impartialité	NOUN	0	24	3.89
6	dualité	NOUN	0	22	3.81
7	poignant	ADJ	1	64	3.76
8	compromettant	VERB	0	21	3.76
9	réévaluation	NOUN	0	21	3.76
10	palpable	ADJ	3	129	3.61

Table 5: Top 10 French AI-overused content words by LPR. c_H and c_M denote human and model counts, respectively, measured in 12-token windows over 100,000 continuations.

E.2 Selected All-Word Lists

Tables 6 and 7 show the top-10 English items by LPR and by volume shift (all POS categories).

#	Lemma	POS	c_H	c_M	LPR
1	additionally	ADV	0	421	6.74
2	emphasize	VERB	16	5 180	5.75
3	revolutionize	VERB	0	59	4.78
4	revitalize	VERB	0	44	4.49
5	renforcer	X	0	36	4.29
6	captivated	VERB	1	94	4.14
7	conservationist	NOUN	0	31	4.14
8	streamlin[e]	VERB	0	27	4.01
9	firsthand	ADV	0	25	3.93
10	personalize	VERB	1	74	3.91

Table 6: Top 10 English AI-overused items by LPR (all categories). c_H and c_M denote human and model counts, respectively, measured in 12-token windows over 100,000 continuations.

#	Lemma	POS	c_H	c_M	Vol. shift
1	and	CCONJ	32 242	41 456	9 214
2	the	DET	47 458	54 692	7 234
3	this	DET	2 673	8 419	5 746
4	emphasize	VERB	16	5 180	5 164
5	that	SCONJ	4 149	8 254	4 105
6	have	AUX	7 710	10 892	3 182
7	importance	NOUN	68	2 684	2 616
8	authority	NOUN	230	2 818	2 588
9	aim	VERB	112	2 664	2 552
10	health	NOUN	797	3 278	2 481

Table 7: Top 10 English items by (rescaled) volume shift ($c_M - c_H$). c_H and c_M denote human and model counts, respectively, measured in 12-token windows over 100,000 continuations. For ease of interpretation, we give a linear rescaling of the absolute prevalence difference $\ell_M(w) - \ell_H(w)$ introduced in Section 3.4.

Additional lists for all languages are available in our anonymous repository (cf. Appendix A).