Fully Automated Identification of Lexical Alignment and Preference-Stage Shifts in Large Language Models

Thomas Stephan Juzek, Xiaoyang Ming, Jose A. Hernandez
Florida State University
{tjuzek, xm24b, jah22q}@fsu.edu

Abstract

The language used by digital chat assistants such as ChatGPT can diverge from human expectations (misalignment). Research, mostly on Scientific English, has described both *what* divergences occur and, to some extent, *why*, linking them to the training stage of human preference learning. Yet, existing approaches rely on manual curation. This paper introduces two curation-free, assumption-light evaluation metrics: the Lexical Alignment Score, which identifies lexical overuse, and the Triangulated Preference Shift, which quantifies how much of such shifts can be attributed to human preference learning. Using PubMed abstracts, continuations were generated and measured using windowed document prevalence across six model families (Falcon, Gemma, Llama, Mistral, OLMo, Yi). The procedure identifies, without manual intervention, overused items such as *suggest*, *additionally*, and *strategy*, and estimates their link to preference learning. Our findings replicate prior work and remain stable across parameter settings, random seeds, and evaluation on further data. The approach scales readily and enables systematic study of lexical (mis)alignment beyond Scientific English and across languages, and as such, the metrics have the potential to contribute to improved alignment for future models and understanding of its origins.

Keywords: lexical alignment, evaluation, Large Language Models, preference learning, Scientific English

1. Introduction

Usage of digital assistants based on Large Language Models (LLMs), such as ChatGPT, is increasing fast, and these artificial intelligence (AI) tools are now widely used for programming, language editing, and information finding (Stack Overflow, 2024; Coffey, 2024; O'Brien and Sanders, 2025; Sidoti and McClain, 2025). They perform strongly on standard benchmarks (e.g., legal, mathematical, and language tasks; Achiam et al., 2023; Hendrycks et al., 2020; Cobbe et al., 2021). Yet they can deviate from human usage in systematic ways. A notable instance of such *misalignment* concerns lexical behaviour: in Scientific English, assistants disproportionately favour items such as "delve", "intricate", and "furthermore" (Matsui, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024; Gray, 2024; Geng and Trotta, 2024). Current evidence implicates the preference-learning stage – e.g., reinforcement learning from human feedback (RLHF; Ouyang et al., 2022) or direct preference optmization (DPO; Rafailov et al., 2024) – as a notable driver of these shifts (Bharadwaj et al., 2025; Juzek and Ward, 2025).

Training of LLM-based chat assistants typically follows four stages: pre-training (giving base models), instruction tuning (to make models more helpful assistants), preference learning (aligning with human judgements), and task-specific fine-tuning (Christiano et al., 2017; Wei et al., 2021; Touvron et al., 2023). For models after instruction and preference tuning, we use *instruct* models throughout the paper; they are sometimes referred to as *chat* models. Importantly, preference learning has delivered substantial gains in assistant behaviour (Ouyang et al., 2022; Rafailov et al., 2024).

However, evaluations of AI-associated lexical items, to both the *what* (model-human misalignment; e.g. Matsui, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024; Geng and Trotta, 2024) and the *why* (stage-specific shifts; e.g. Juzek and Ward, 2025), face a major gap: dominant approaches rely on manual curation and filtering (mainly ad hoc heuristics), which limits reproducibility and scalability across domains and languages.

We introduce a curation-free evaluation pipeline over model generations. We treat lexical overuse as a tractable behavioural probe, and the main contribution is a diagnostic within the alignment pipeline, with pointers to attribution (causes for the observed model behaviour). For this, we evaluate six model families, with both base and instruct variants, using 42,000 PubMed abstracts as input. All generations use deterministic decoding, and symmetric cleaning is applied to human and model texts alike (Section 3). We then introduce two metrics: the *Lexical Alignment Score* (LAS), which quantifies *what* is overused relative to human continuations (Section 4.3); and the *Triangulated Preference Shift* (TPS), which isolates the *why*, that is, uplifts attributable to the preference-learning stage (Section 4.4). The metrics give promising results, both for individual lexical items and for estimating

Code: github.com/fsu-nlp/lexical-alignment-shifts. Correspondence: TSJ. Contributions: TSJ led conceptualisation/methodology, implementation, validation/analysis, and writing, assisted by XM and JH.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6116–6131
11–16 May 2026.

how much of the observed shifts stem from preference learning, as well as macro-level model trends (Section 5). We validate the metrics by analysing additional data, varying parameters, and comparing results against prior literature (Section 6). Our work, by providing more effective diagnostics, lays the foundations for more effective mitigation, which remains to be developed further. This situates our study alongside recent work suggesting that post-training can improve assistant behaviour while also narrowing output diversity, collapsing preference variation, or amplifying dominant response patterns (Kirk et al., 2023; Xiao et al., 2024; Zhang et al., 2025; Murthy et al., 2025). Broader implications are discussed in Section 7.

2. Related Work

2.1. Lexical Overuse in LLMs

After the release of ChatGPT, a sudden spike in the usage of a small set of words (e.g., "delve", "furthermore", "intricate") in academic writing was noted (Matsui, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024). As many of these words are also overused in AI-generated texts (Matsui, 2024), the most plausible explanation for the spikes is that AI-assisted writing has contributed to these spikes (Kobak et al., 2024). Whilst most work concentrates on Scientific English, there are parallel findings in journalism/newsroom contexts (Fitterer et al., 2025). These analyses are largely based on written corpora; some work notes the rise of AI-associated language in unscripted speech (Yakura et al., 2024; Anderson et al., 2025), but causal attribution remains incomplete.

2.2. Mechanisms Underlying Lexical Overuse

Evidence implicates RLHF/DPO as *a* driver of lexical overuse in Large Language Models. Triplet comparisons (Human vs Base vs Instruct) show patterns consistent with the hypothesis that preference learning plays a role (Bharadwaj et al., 2025; Juzek and Ward, 2025). For stylistic and formatting choices, small biases in preference data can produce disproportionately large behavioural changes (Zhang et al., 2024). Work on self-training and synthetic instruction-tuning data highlights risks of feedback loops and style drift, which include possible output degradation (Alemohammad et al., 2023; Briesch et al., 2023; Shumailov et al., 2023). These findings further motivate automated diagnostics that can separate general misalignment from preference-stage effects.

2.3. Limitations of Prior Identification Procedures

Most procedures proposed in the literature to identify AI-overused words have two shortcomings (studies commonly show one or both; representative examples include Kobak et al., 2024; Gray, 2024; Matsui, 2024; Juzek and Ward, 2025): First, they rely on manual curation, particularly hand-tuned heuristics and post hoc filtering, which limits its scalability, reproducibility, and cross-domain and cross-language applications. The discourse is mostly English-centred, with cross-language exceptions in German, Spanish, Portuguese, Chinese, and Japanese, largely, however, in the area of AI detection and stylometric; (Irrgang et al., 2024; da Silva and Rottava, 2024; Kotz et al., 2024; Jin et al., 2025; Zaitsu and Jin, 2023; Schaaff et al., 2024; Terčon and Dobrovoljc, 2025). Second, they often rely on data for which the production process is unclear, and it is then *assumed* that sudden changes in writing are due to AI. Concretely, many papers analyse scientific abstracts before and after 2022, note the stark rise of items like "delve", and attribute the rise to AI. This is a plausible inference, but it remains an assumption.

2.4. Alignment, Misalignment, and Possible Influences on Humans

The general goal of alignment is to make systems reflect human goals, values, beliefs, and broader expectations (Russell et al., 2015; Gabriel, 2020). Alignment research often proceeds through a sequence of diagnostics, characterisation, attribution, and mitigation. In digital assistants, alignment is commonly operationalised via preference learning (RLHF/DPO) (Ouyang et al., 2022; Rafailov et al., 2024); recent work operationalises value alignment measurements for LLMs (Norhashim and Hahn, 2024). However, the literature documents wider misalignments and societal harms (Blodgett et al., 2020; Bender et al., 2021), as well as more specific issues, such as gender and race biases in LLM behaviour and downstream domains (Kotek et al., 2023; Omiye et al., 2023), and evaluation artefacts that confound competence with stylistic preferences (Perez et al., 2023).

A more benign form of misalignment is sycophancy, where AI assistants compliment users irrespective of truth; there are efforts to identify and mitigate sycophancy (Sharma et al., 2023; Wei et al., 2025); similarly for verbosity (Park et al., 2024). High-stakes value-level misalignment is exemplified by DeepSeek-R1 (Gibney, 2025; Allen, 2025): on topics central to liberal-democratic discourse, responses were reported to be censored or heavily reframed towards narratives aligned with political entities (Samuel and Crook, 2025; Stokel-

Walker, 2025; Naseh et al., 2025; Cole, 2025). Such model behaviour is at odds with widely endorsed free-expression norms in liberal-democratic contexts (Poushter et al., 2025). The societal risk is an undesired normative drift: repeated, low-salience exposure could normalise such narratives (mere-exposure/illusory-truth effects; Zajonc, 1968; Hasher et al., 1977), and LLM-authored political messaging might shift attitudes (Bai et al., 2025).

Our work is concerned with lexical choice alignment, and in particular the steps of diagnostics, characterisation, and attribution, with pointers to mitigation. The overarching theme is that of alignment side effects, and our work is informed by the discourse around value and content alignment. Thus, structural findings on lexical (mis)alignment, particularly concerning the role of preference learning, could connect to issues of value and content (mis)alignment and could be of value for those discourses.

2.5. Relation to AI detection

In addition, a large literature explores AI detection (classifier-based, watermarking, perplexity/stylometry) and a large body of domain-specific corpora for detection exists (inter alia, Gehrmann et al., 2019; Chakraborty et al., 2023; Mitchell et al., 2023; Kirchenbauer et al., 2024; Huang et al., 2025). Strengths of AI detection include available automation and analysis of multiple linguistic and extra-linguistic levels; limitations include high false-positive rates and relatively low robustness across domains and languages (Sadasivan et al., 2023; Weber-Wulff et al., 2023). Detection work addresses attribution, not linguistic-level-specific behaviour and stage-specific mechanisms. While both detection and diagnostics/identification approaches, such as our work, share the goal of automated, scalable pipelines, their overarching aims differ (detection: classification of instances; diagnostics/identification: identifying macro-level characteristics of model behaviour).

3. Task and Data

We address the gaps identified in Section 2 by introducing curation-free, assumption-light metrics that cover both the *what* and *why* of lexical (mis)alignment. Our data comprise PubMed abstracts, which enables a comparison with the literature on lexical overuse in LLMs, which largely focuses on this very domain (Scientific English). We sample 42,000 abstracts from 2012-2021, i.e., the ten years preceding the release of ChatGPT. Per year, 4,200 abstracts were, without replacement, sampled. Each model (as per Table 1) generated the second part of an abstract, totalling a data basis of 63.4 million lemmatised tokens. Due to the windowing approach described in Section 4, with a window size of 50 lemmatised tokens, this set-up provides coverage of approximately 2m lemmatised tokens per model variant after accounting for exclusions. Each abstract was split at the sentence boundary closest to its midpoint using a Python script. The first halves served as prompts for model generation; the second halves constitute the human gold standard. This design is related to cloze-style and human-continuation evaluation, where model predictions are assessed against paired human responses (Eisape et al., 2020; Giulianelli et al., 2023; Ilia and Aziz, 2024). Because sentence boundary detection is not fully error-free, we manually spot-checked 100 abstracts. In 96 cases, the split was at a true sentence boundary; in 4 cases, the splits were non-ideal. We treat split variance as a minor source of preprocessing error, unlikely to materially affect the downstream results.

3.1. Models and Decoding Policy

For each prompt index $r \in \{1, \ldots, R\}$, we generate one continuation per model. The basic computational set-up is described in Section 8. We used deterministic greedy decoding throughout. Sampling was disabled (temperature fixed at 0), and nucleus and top-k controls were turned off. Decoding terminated on the model's end-of-sequence token (<eos>) or at the model token limit T. Through <eos> suppression, we enforced a minimum of 120 tokens, and we capped outputs at 200 tokens. To avoid generation loops, we enforced a 4-gram no-repeat constraint. Variation was further minimised by fixing a global seed. The pad token was set equal to <eos>. These settings are consistent across all generations.

Base models are next-token generators and simply received the prompt r_i as input for plain next-token continuation. For instruct models, we used a minimal wrapper: a system message ("Reply only with the continuation; do not repeat the user text; no preface.") and a user message containing the first half of an abstract r_i were passed. Loop/meta removal, safety messages, and meta-chat were handled symmetrically during the cleaning stage.

3.2. Model Families

Model families were selected to (i) provide both base and instruct variants and (ii) support temperature $= 0$ decoding to maximise reproducibility. Under these criteria, we chose six popular families: Falcon (Almazrouei et al., 2023), Gemma (Gemma Team, 2025), Llama-3.1 (Grattafiori et al., 2024), Mistral-v0.3 (Jiang et al., 2023), OLMo-2 (Team OLMo et al., 2025), and Yi-1.5 (Young et al., 2024). We used the mid-sized models; in these families,

Family	Params	Type	Repository	Revision (sha7, date)
Falcon-3	7B	Base	`tiiuae/Falcon3-7B-Base`	`bf3d7ed` 24-12-17
		Instruct	`tiiuae/Falcon3-7B-Instruct`	`1e57a0e` 25-05-31
Gemma-3	4B	Base	`google/gemma-3-4b-pt`	`cc012e0` 25-03-21
		Instruct	`google/gemma-3-4b-it`	`093f9f3` 25-03-21
Llama-3.1	8B	Base	`meta-llama/Llama-3.1-8B`	`d04e592` 24-10-16
		Instruct	`meta-llama/Llama-3.1-8B-Instruct`	`0e9e39f` 24-09-25
Mistral	7B	Base	`mistralai/Mistral-7B-v0.3`	`caa1feb` 25-07-24
		Instruct	`mistralai/Mistral-7B-Instruct-v0.3`	`0d4b76e` 25-07-24
OLMo-2	7B	Base	`allenai/OLMo-2-1124-7B`	`7df9a82` 25-01-06
		Instruct	`allenai/OLMo-2-1124-7B-Instruct`	`470b1fb` 25-01-06
Yi-1.5	6B	Base	`01-ai/Yi-1.5-6B`	`157a3d7` 24-06-26
		Instruct	`01-ai/Yi-1.5-6B-Chat`	`771924d` 24-08-27

Table 1: An overview of the models used, with model size in parameters, model types, repositories, and pinned revisions. Short SHAs (last 7 characters) and release dates are given.

larger variants typically add multimodality with limited improvement on purely textual tasks (Hugging Face Team, 2024). This gives the base-instruct pairs listed in Table 1.

3.3. Cleaning and Part-of-Speech Tagging

Cleaning was deletion-only and symmetric (to avoid measuring pre-processing artefacts), i.e., applied to the human gold standard as well as to both base and instruct generations. Cleaning proceeded in two stages. First, we applied a deterministic regex pre-clean: normalise whitespace; collapse runs and newlines; strip leading and trailing spaces; and drop any text following the most frequent signal that the abstract has concluded and the model is continuing into the article body ("Introduction"). All removals were logged. Second, for information that is difficult to remove reliably with pattern matching, we used GPT-4.1-mini with temperature set to 0 to delete AI persona and meta text ("Certainly, here is ..."), dialogue scaffolding ("<assistant>"), loops (keeping one copy), and first- and second-person material ("Can you explain the meaning of ..."). We instructed the cleaner to preserve genuine abstract content and not to paraphrase or reorder (outputs include diffs and run summaries). The exact GPT-4.1 system prompt follows best practices and can be found in Appendix A. The cleaning prompt was simply:

```
"Clean the following MID-ABSTRACT
    CONTINUATION by applying ONLY the
    deletion rules."
"Do not paraphrase or rewrite; return
    the cleaned text only.\n\n"
f"INPUT:\n{raw}\n\n"
"OUTPUT (cleaned text only):"
```

To analyse different inflected forms (e.g., 'delve', 'delved', 'delving') under a single lemma type, and to distinguish words that share a surface form but differ in meaning/usage (e.g., 'analysed' as a verb vs an adjective), we tag the cleaned generations for part-of-speech (POS). This also enables analyses of additional categories (e.g., excluding certain part-of-speech categories from certain computational steps). We used spaCy 3.8 (`en_core_web_trf`; Honnibal et al., 2020) for tagging, with the Universal Part-of-Speech (UPOS) tags from the Universal Dependencies framework (Nivre et al., 2020); outputs are in CoNLL–U format (Zeman et al., 2017).

4. Evaluation Metrics

Both metrics introduced below, the *Lexical Alignment Score* (LAS) and the *Triangulated Preference Shift* (TPS), operate on length-controlled windows of paired human-model continuations, using windowed document prevalence to ensure robustness against few-document spikes.

Intuitively, the two metrics approach the following questions. LAS asks whether a model uses a given lexical item more or less often than humans do under the same prompt. TPS asks whether such overuse appears to arise mainly after the base stage, rather than already being present in the base model. In this sense, LAS targets the *what* of lexical divergence, whereas TPS targets part of the *why*.

4.1. Windowed Document Prevalence

Per abstract, the first half serves as the prompt; the second half is the human gold standard H. Using the prompt, we generate continuations for the base model (B) and instruct model (I).

Plain relative frequencies are fragile: as corpus size grows, the probability of distortions rises whenever a plausible term occurs repeatedly in only one *stream* S (human/base/instruct); say e.g., two prompts elicit multiple "EuroQol" mentions in I but not in H or B. A reasonable fix for this is *document frequency* (did the item occur at all?); however, models may systematically produce docu-

ments of different lengths, giving items varying degrees of opportunity to occur. We therefore score presence within a fixed-size window, $K{=}50$ lemmatised tokens by default (the code allows variation), placed at a random region of the document (with a deterministic seed for the purposes of this work). Concretely, we choose a percentile offset $\pi_r \in (0, 1)$ deterministically for the prompt ID and use the same π_r across H-B-I for that prompt, in order to ensure paired/triplet comparability (pairs for LAS; triplets for TPS). Pre-processing is symmetric across $H/B/I$ (see Section refsec:cleaning); if any continuation is too short after cleaning, the entire pair/triplet is removed. This gives a length-controlled, spike-robust prevalence signal. An informal example of the window approach (using words instead of lemmatised tokens) is as follows. Suppose the entire document is "This is a sentence with seven words", and the window size is $K{=}3$. The sentence has 7 words, so there are $7{-}3{+}1 = 5$ possible 3-word windows. A random number between 0 and 1 is drawn, say 0.70 (the 70th percentile). Multiplying this by 5 and rounding down gives a starting index of 3 (counting from 0). This corresponds to positions $4 - 6$ when counting from the first word. Taking these three words gives the window: "This is a [sentence with seven] words".

Formal Definition. We write $S \in \{H, B, I\}$ for the stream and index prompts by $r = 1, \ldots, R$. Each pair (r, S) constitutes a document $D_{r,S}$ (the human second half if $S{=}H$, or the model continuation if $S \in \{B, I\}$). A stream S is the collection $\{D_{r,S} : r = 1, \ldots, R\}$. Let w denote a lemma type with UPOS tag, and let $\omega(\cdot)$ be the lemmatisation map from lemmatised tokens to lemma types; for a token t, $\omega(t)$ is its lemma+UPOS type. For prompt $r \in \{1, \ldots, R\}$ and stream S, we define:

$$Y_{rS}(w) = \mathbf{1}\Big\{ w \text{ appears in the window of size } K\text{-lemmatised tokens of } S \text{ at perc. } \pi_r \Big\}$$

$$c_S(w) = \sum_{r=1}^{R} Y_{rS}(w)$$

Intuitively, $Y_{rS}(w)$ is a per-document window flag: 1 if w occurs at least once in the K-lemmatised-token window of document $D_{r,S}$, else 0. Summing over prompts gives $c_S(w) = \sum_r Y_{rS}(w)$, the windowed document frequency for stream S. For example, if "EuroQol" appears many times in the I-window for two prompts only, then $c_I(\text{EuroQol}) = 2$. Here R is the number of retained prompts after symmetric cleaning, so $c_S(w) \in \{0, 1, \ldots, R\}$. To avoid degenerate $0/1$ rates (infinite log-odds) and reduce small-sample variance for rare lemmas, we estimate window-prevalence with Jeffreys smoothing (a common smoothing method for document-frequency estimation, Krichevsky and Trofimov, 1981), which is essentially adding a half "pseudo-hit" and a half "pseudo-miss" to stabilise the estimate, with Beta$(\frac{1}{2}, \frac{1}{2})$ prior:

$$\ell_S(w) = \frac{c_S(w) + \frac{1}{2}}{R + 1}$$

where $\ell_S(w)$ is the Jeffreys-smoothed *windowed prevalence* of w in stream S.

4.2. Two Stages

Both LAS and TPS run in two consecutive stages.
Estimation. On the cleaned $H/B/I$ documents, we compute $\ell_S(w)$ for *all* UPOS-tagged lemma-types (incl. PUNCT). UPOS usage (e.g., punctuation rates) may differ between human and model texts; filtering here could induce selection effects and artefactual convergence/divergence even for non-filtered categories. This gives us windowed document-prevalence estimates for each stream-lemma pair (S, w), which are the basis for both LAS and TPS. Programmatically, this step builds a look-up table from lemma-types (lemma+UPOS) to (i) $\ell_S(w)$ and to (ii) derived per-lemma-type contrasts used by the two metrics, LAS and TPS.
Scoring. For both LAS and TPS, we use the estimated $\ell_S(w)$ to score lemmatised tokens, sequences, documents, and corpora/models; essentially, using the look-up table for scoring. We report on the full tag set; the code allows tag selections.

4.3. Lexical Alignment Score

The Lexical Alignment Score quantifies the alignment of a model M (either B or I) relative to H (while one could compare B and I directly, H is the natural reference; for three-way contrasts, the Triangulated Preference Shift metric is more useful). Estimation is performed at the lemma+UPOS level and later used for scoring lemmatised tokens, sequences, documents, and corpora/models.

Estimation. Estimation is per UPOS-tagged lemma-type, using the paired prompt-aligned windows defined in Section 4.1. For a lemma type w and a model $M \in \{B, I\}$ we define the lemma-type-level LAS:

$$\mathrm{wLAS}_M(w) \;=\; \ell_M(w) \;-\; \ell_H(w)$$

where $\ell_S(w)$ is the Jeffreys-smoothed windowed prevalence in stream S. Positive values indicate overuse relative to human usage; negative values indicate underuse. The following example illustrates this: Suppose that a lemma occurs in 18% of model windows but in 10% of human windows, and is thus overused. Then its word-level LAS is $0.18 - 0.10 = 0.08$. If the values were instead 7% versus 10%, thus underuse, the LAS score would be -0.03.

Rank	lemma+UPOS	wLAS
1	these_DET	0.162
2	to_PART	0.123
3	suggest_VERB	0.093
4	finding_NOUN	0.080
5	this_DET	0.078
6	the_DET	0.078
7	a_DET	0.077
8	furthermore_ADV	0.077
9	such_ADJ	0.076
10	research_NOUN	0.076
11	additionally_ADV	0.074
12	to_ADP	0.070
13	for_ADP	0.070
14	further_ADJ	0.069
15	study_NOUN	0.069
16	that_SCONJ	0.065
17	,_PUNCT	0.061
18	could_AUX	0.061
19	into_ADP	0.060
20	highlight_VERB	0.059

(a) Top-20 wLAS entries, aggregated over all instruct models.

Rank	lemma+UPOS	wLAS
3	suggest_VERB	0.093
4	finding_NOUN	0.080
8	furthermore_ADV	0.077
9	such_ADJ	0.076
10	research_NOUN	0.076
11	additionally_ADV	0.074
14	further_ADJ	0.069
15	study_NOUN	0.069
20	highlight_VERB	0.059
21	potential_ADJ	0.055
22	include_VERB	0.053
24	lead_VERB	0.053
25	crucial_ADJ	0.053
28	understand_VERB	0.049
30	role_NOUN	0.048
31	enhance_VERB	0.047
32	involve_VERB	0.045
34	researcher_NOUN	0.045
35	various_ADJ	0.044
36	strategy_NOUN	0.044

(b) Top-20 wLAS entries for nouns, verbs, adjectives, adverbs, aggregated over instruct models.

Table 2: Top lexical shifts by lemma-type (wLAS). Higher values indicate model overuse vs human.

Scoring. Given these per-lemma-type scores $\mathrm{wLAS}_M(w)$, we define per-lemmatised-token contributions $\Delta(t)=\mathrm{wLAS}_M(\omega(t))$ for lemmatised token t in a unit U (lemmatised tokens, sequence, document, or corpus) with lemmatised-token multiset $\mathcal{T}(U)$. We score U by the *L2 mean* (Bishop and Nasrabadi, 2006):

$$\mathrm{uLAS}(U;M)=\left(\frac{1}{|\mathcal{T}(U)|}\sum_{t\in\mathcal{T}(U)}\Delta(t)^2\right)^{1/2}$$

This gives a length-normalised root-mean-square (RMS) magnitude that avoids cancellation between over- and under-use, while placing greater emphasis on large deviations. Aliases for readability: $\mathrm{sLAS}(s;M)=\mathrm{uLAS}(s;M)$ (sequence), $\mathrm{dLAS}(d;M)=\mathrm{uLAS}(d;M)$ (document), and $\mathrm{cLAS}(C;M)=\mathrm{uLAS}(C;M)$ (corpus).

Model Aggregation. Scores may be aggregated across models; we take a simple macro-average, giving equal weight to every model irrespective of the number of documents/lemmatised tokens analysed. This is most informative at the lemma type level, which gives an overview of the most misaligned lemmas across LLMs. Section 5 reports model-aggregated wLAS and the strongest shifts.

Scoring Example. s_1 and s_2 illustrate sequence-level scoring (casing/lemmatisation/cleaning omitted). Numbers in the second row are the per-lemmatised-token contributions $\mathrm{wLAS}_M(\omega(t))$. uLAS is the length-normalised L2 mean.

s_1:	Our	work	shows	how	⋯
$\Delta(t)$	-0.027	-0.005	-0.002	0.001	⋯

s_2:	Our	work	highlights	how
$\Delta(t)$	-0.027	-0.005	0.027	0.001

We score each sequence with the $L2$ mean: square $\Delta(t)$ values, average over the lemmatised token count ($|\mathcal{T}(s_i)|=4$), then take the square root, resulting in $\mathrm{sLAS}(s_1)\approx 0.014$ and $\mathrm{sLAS}(s_2)\approx 0.019$.

4.4. Triangulated Preference Shift

The Triangulated Preference Shift metric isolates preference-stage effects by requiring an instruction-tuned model I to exceed both the human baseline H and the base model B, while penalising shifts already present in B. Estimation is performed at the lemma+UPOS level and then used for scoring different linguistic units.

Estimation. Estimation is done for lemma-types w and a model family M with streams B_M (base), I_M (instruct), and a common human baseline H. With $\ell_S(w)$ being the Jeffreys-smoothed windowed prevalence in stream S (Section 4.1), we get:

$$\Delta_{IH}(w)=\ell_I(w)-\ell_H(w)$$
$$\Delta_{IB}(w)=\ell_I(w)-\ell_B(w)$$
$$\Delta_{BH}(w)=\ell_B(w)-\ell_H(w)$$

The lemma-type-level triangulated pref. shift is:

$$\mathrm{wTPS}_M(w) = \min\{\Delta_{IH}^{(M)}(w),\, \Delta_{IB}^{(M)}(w)\} - \max\{0,\, \Delta_{BH}^{(M)}(w)\}$$

Rank	lemma+UPOS	wTPS
1	to_PART	0.130
2	these_DET	0.112
3	this_DET	0.085
4	to_ADP	0.076
5	furthermore_ADV	0.075
6	such_ADJ	0.073
7	research_NOUN	0.070
8	for_ADP	0.069
9	additionally_ADV	0.065
10	further_ADJ	0.064
11	study_NOUN	0.061
12	,_PUNCT	0.058
13	highlight_VERB	0.057
14	into_ADP	0.057
15	potential_ADJ	0.056
16	finding_NOUN	0.055
17	a_DET	0.054
18	as_ADP	0.054
19	could_AUX	0.053
20	crucial_ADJ	0.052

Table 3: Top 20 overused lemma+UPOS by wTPS.

Informal reading: wTPS measures the portion of the instruct model's uplift that cannot be explained by the base model. It is positive only when I beats both H and B, and by more than any pre-existing $B > H$ lean; otherwise it is zero or negative.

The following example illustrates the score. Suppose that a lemma occurs in 10% of human windows, 12% of base-model windows, and 20% of instruct-model windows. Then $\Delta_{IH} = 0.10$, $\Delta_{IB} = 0.08$, and $\Delta_{BH} = 0.02$. TPS is therefore $\min(0.10, 0.08) - 0.02 = 0.06$, i.e. a positive residual shift beyond what was already present in the base model. By contrast, if the base model were already at 18%, TPS would be much smaller or vanish.

Scoring. Given $\mathrm{TPS}(w)$, we compute an $L2$ root-mean-square (RMS) score uTPS that applies uniformly to sequences, documents, and corpora (similar to uLAS). For any unit U with lemmatised token multiset $\mathcal{T}(U)$ and lemmatisation map $\omega(\cdot)$:

$$\mathrm{uTPS}(U; M) = \left(\frac{1}{|\mathcal{T}(U)|} \sum_{t \in \mathcal{T}(U)} \mathrm{wTPS}_M\big(\omega(t)\big)^2 \right)^{1/2}$$

This gives a length-normalised RMS magnitude that avoids cancellation between over- and underuse, and weights larger deviations more strongly. For reporting, we focus on upward shifts, using $\max(0, \mathrm{wTPS}(w))$. Aliases: $\mathrm{sTPS}(s) = \mathrm{uTPS}(s)$ (sequence), $\mathrm{dTPS}(d) = \mathrm{uTPS}(d)$ (document), and $\mathrm{cTPS}(C) = \mathrm{uTPS}(C)$ (corpus).

Model Aggregation. As with LAS, TPS scores may be aggregated across models.

5. Results

Table 2a reports the top 20 lemma-level Lexical Alignment Scores aggregated over all instruct models; results for the base models can be found in Table 4 in Appendix B. Table 2b links to prior work by limiting analysis to content words (NOUN/VERB/ADJ/ADV) and aggregating over instruct models. The top 20 items ranked by the preference-stage metric are listed in Table 3. Macro-level alignment, corpus-level LAS, by family and stage is shown in Figure 1 (lower values indicate closer alignment to humans). Figure 2 shows the corpus-level TPS *ratio* R, defined as $R = \mathrm{cTPS}_{\text{instruct}}/\mathrm{cTPS}_{\text{base}}$, where higher values of R indicate stronger preference-stage shifts in the instruct variant. The results show that the proposed diagnostics identify lexical items already discussed in the literature and that, although instruct models are often closer to humans overall, this pattern weakens or reverses when attention is restricted to content words.

6. Validation

To further validate the metrics, we conducted four checks. (i) we scored an additional, unseen 20% of the data. (ii) We compare our overuse list to the literature. (iii) We ran the metrics with different window sizes ($K = 40, 50, 60$). (iv) We reran the $K = 50$ configuration for four more random seeds. The following is done for the LAS metric.

Additional Data. An additional 20% of data, comprising 8,400 abstracts (840 per year), was scored. The results per model are (base vs instruct): Falcon: 0.0346 vs 0.0306, Gemma: 0.0332 vs 0.0326, Llama: 0.0331 vs 0.0338, Mistral: 0.0333 vs 0.0320, OLMo: 0.0345 vs 0.0317, Yi: 0.0329 vs 0.0304. The human baseline comes out at 0.0304. The results are extremely similar to those reported in Figure 1.

Convergence with Prior Research. Prior work provides curated inventories of AI-associated words. Geng and Trotta (2024) discuss 8 words in-depth. Galpin et al. (2025) list 32 overused words (surface-form level, i.e., including inflections), Kobak et al. (2024) list 291 such items. We matched these against our overused set (focus: instruct models); for ambiguous surface forms (e.g., 'diminishing', which can occur as a verb or an adjective), we counted a hit if any corresponding lemma+UPOS variant appeared. 8/8 words from Geng and Trotta (2024) feature prominently in our data. From the list in Galpin et al. (2025), 32/32 en-

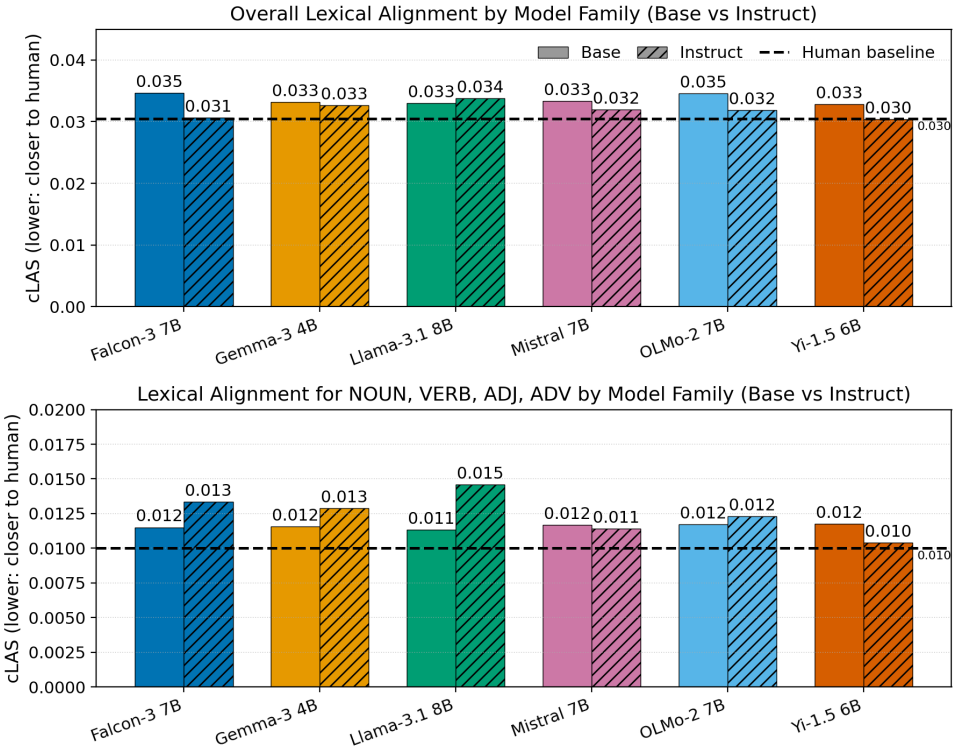

Figure 1: cLAS by model family (base vs instruct); lower: closer to human.

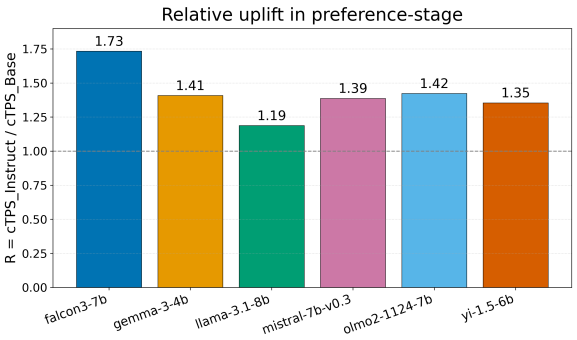

Figure 2: cTPS ratios (base vs instruct).

tries occur in our list. From the 291 items in Kobak et al. (2024), 240 (83%) occur in our data.

Varying K. We observe that scores vary slightly across different window sizes. Here, we choose $K = 40$ (smaller windows disregard a larger portion of the data), 50, and 60 (window sizes larger than that approach document length and undermine the windowing logic). The scores averaged across model families for base vs instruct models (with the human baseline in parentheses) are: 0.357 vs 0.354 (0.324), 0.336 vs 0.319 (0.3048), and 0.324 vs 0.318 (0.305), respectively. The general trend that base > instruct > human remains consistent across window sizes. Some variation is expected, partly due to the changing proportional effect of Jeffreys smoothing as window size increases.

Different Seeds. The global average over all base and instruct models at seed 42 is 0.03275. For four additional seeds, the results are 0.03276 (s=43), 0.03277 (s=44), 0.03283 (s=45), and 0.03281 (s=46). The outcomes are very robust.

Further Validation. The TPS could be further validated by analysing instruction and preference datasets to assess whether the metric aligns with the linguistic patterns those datasets encode. The datasets used in OLMo could be particularly well suited for this purpose, as Allen AI have made the entire process, including all data, publicly available.

Additional validation could also come from controlled model development: deliberately inducing controlled doses of a model bias and testing whether it manifests in model behaviour. This would parallel Zhang et al. (2024), who analysed style- and format-related model behaviours.

7. Discussion

At the corpus level, instruct variants generally align *better* with human usage than their base counterparts (5/6 families; Figure 1 top). However, focusing on content words (NOUN/VERB/ADJ/ADV), for all models this trend either reverses (4/6 families in Figure 1 bottom) or at least weakens. These are precisely the items emphasised in prior reports of lexical overuse. However, the largest aggregate shifts are dominated by function markers (Table 2a),

consistent with the findings that AI models have their own syntax style (Zamaraeva et al., 2025). The Triangulated Preference Shift metric corroborates stage influence: many high-LAS items also show a positive TPS, indicating uplift in instruct over both human and base (Table 3), which affects some model families more than others (Figure 2). Per-model results are feasible and important for model-individual audits, and per-model results can be found in Appendix C.

An important observation is that quite a few of the top-ranked items are not unusual scientific terms. We interpret this as evidence that the models amplify already typical scientific wording, pushing high-prevalence items toward even greater dominance. This observation also highlights an aspect of the metrics: they are most sensitive to prevalence *volume* and, arguably, though this remains to be explored, less sensitive to low-frequency words that show small absolute but large relative shifts ("spikes"). Such effects might be captured more directly by ratio-based measures, for example using log prevalence ratios between human and model outputs. At the same time, even these lower-volume but higher-relative shifts may reflect a related tendency, namely the amplification of lexical patterns that are already present, at least in emerging form, in human writing. We regard both phenomena as important, and therefore emphasise the value of systematic comparisons across human texts, base models, and instruction-tuned models.

Our work is situated within the discourse on lexical overuse. In principle, the Lexical Alignment Score captures both overuse and underuse. The latter is generally less noteworthy, as it tends to show less disruptive shifts (Galpin et al., 2025). By contrast, the Triangulated Preference Shift metric measures foremost positive shifts that are attributable to preference learning.

A broader relevance of our work stems from the observation that AI-associated lexicon is now attested in spontaneous speech (Yakura et al., 2024; Anderson et al., 2025), indicating rising prevalence in human language use. While our work does not establish adoption directly, our findings identify model-side lexical biases that are plausible candidates for such uptake. AI chatbots may be accelerating, and possibly causing, these shifts; therefore, the ability to align models with human linguistic expectations is valuable. One plausible pathway is repeated exposure during AI-assisted writing and interaction: psycholinguistic work shows that speakers tend to align with recently encountered linguistic forms, including those produced by computer and robot interlocutors (Brennan, 1991; Branigan et al., 2003; Brandstetter et al., 2017; Ostrand et al., 2023). Work on AI-mediated communication further suggests that such exposure can reduce diversity in co-writing (Padmakumar and He, 2023) and may, in some settings, leave persistent linguistic traces in subsequent human communication (Riedl et al., 2024). More broadly, our work contributes to diagnostics, characterisation, and partial attribution, mirroring core stages of alignment research in other domains. Although we do not study value alignment directly, the present findings help illuminate how preference-based post-training can generate measurable behavioural side effects, and may therefore inform wider mechanism-focused discussions of model alignment.

A question that lies beyond the scope of this paper, but is important for the field more broadly, is whether the underlying phenomenon truly constitutes *mis*alignment. The alignment procedures themselves may be effective: models might indeed align to the preferences present in the preference, learning datasets; preferences expressed by data workers, but these may differ from the expectations of end users. For a discussion of this issue, see Santurkar et al. (2023) and He et al. (2024).

8. Conclusion

The goal of this paper was to introduce evaluation metrics for measuring lexical alignment and the contribution of preference learning. Critically, the proposed methods are curation-free and assumption-light. Key to it is a scalable design in which model continuations are compared against matched human continuations from the same source documents.

The Lexical Alignment Score and Triangulated Preference Shift offer promising approaches for quantifying both the *what* (how closely does model usage match human usage) and the *why* of lexical alignment (how much divergence is attributable to the instruction/preference stage), with clear potential to inform model development. Importantly, these metrics also enable the study of lexical (mis)alignment beyond Scientific English and across languages other than English. This matters because LLM-based chat assistants are now used by millions of people worldwide, who are continually exposed to their linguistic outputs. Ensuring that these outputs align with human expectations is therefore of growing societal importance.

Acknowledgements

We thank Gordon Erlebacher and Zina Ward for their valuable input, and the reviewers for their very constructive feedback. We are grateful to the Florida State University Research Computing Center for computational support.

Code, Data, Computing Set-up

Code and Data. All code, with notes on how to retrieve data, is available at: github.com/fsu-nlp/lexical-alignment-shifts.

Computational Set-up. All major computations were run on two machines.

(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.

(B) University HPC node. NVIDIA RTX A4500 (20 GB). AMD EPYC 7313 (16 cores); 251 GiB RAM. Linux 4.18.0-372.32.1.

Software (identical on both). 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; `spaCy` 3.8.7.

Ethical Considerations

This work presents minimal risks: data are public PubMed abstracts and model outputs; procedures are open and aimed at reducing bias and improving fairness; misuse potential is low; and compute was reasonable (six models; cost about $1160). A broader ethical concern, however, is the labour behind preference-learning datasets, which is often precarious (Perrigo, 2025, 2023). While not inherent to the technology, this is part of its current realisation and warrants continued scrutiny.

Limitations

Our work has several limitations, of which we discuss the most notable ones. Firstly, the present design is limited in that the human gold standard always comes from the second half of the abstract. This may introduce position effects: some lexical items are more likely to occur early in abstracts, whereas others are more likely to occur later. For instance, markers such as *firstly* may be associated with earlier text positions, while items such as *lastly* may be associated with later ones. Some divergences may therefore partly reflect document-position asymmetries rather than model behaviour alone. Future work should extend the design to windows or gold-standard continuations drawn from all text positions.

Second, it is restricted to Scientific English; and more precisely, to the language of continuations of Scientific English abstracts, as abstract-initial language will most likely be scarce in our data. This restriction is motivated both by cost considerations (running six models within a single domain well exceeded $1,000) and by the need to ensure comparability with the existing literature. That said, the project is expressly motivated by scaling this line of work beyond Scientific English.

Third, we do not include the most popular chat assistant, ChatGPT, because since GPT-4, there is no base variant public, which prevents triangulation, and because since GPT-5, temperature cannot be set to 0 (it has a temperature parameter, it just cannot be set to 0), which hampers reproducibility.

Further, there is a very high likelihood of some overlap between our PubMed-based evaluation set and model pre-training data. This may raise absolute human–model similarity, especially for base models; however, it is less likely to account for within-family base-instruction contrasts, which primarily reflect post-pre-training changes.

Appendix A: System Prompt Cleaning

For the symmetric cleaning step, generated and human continuations were processed with the following system prompt:

```
ROLE: Editorial cleaner for SCI-
ENTIFIC ABSTRACT CONTINUATIONS
(mid-abstract).
ACTIONS: ONLY DELETE text; NEVER
paraphrase, reorder, merge, or
add.  Keep sentence wording and
order.  If uncertain whether to
delete, KEEP.


DELETE:
1) Meta/AI persona:  e.g., "Cer-
tainly, here is ...", "as an AI
model", apologies, instructions,
tool/safety notes.
2) Conversation turns & scaf-
folding (only if truly dia-
logic):  pseudo-dialogue mark-
ers <|user|>, <|assistant|>,
</|user|>, </|assistant|> ONLY
IF followed by chat-like material
(greeting, instruction, apology,
question to reader); delete the
marker and that span; otherwise
delete markers only.
3) Obvious repetition/loops:  re-
move verbatim or near-verbatim
repeats; KEEP one copy.
4) First/second-person META sen-
tences using "I/me/my" or di-
rect address "you/your".  Do NOT
delete "we/our/us".


PRESERVE:
- Keep phrases like "In conclu-
sion" / "concluding" when embed-
ded in a normal sentence.
- All scientific content and
phrasing (incl. "we/our/us").
```

```
- Angle-bracketed tokens in gen-
eral (operators, tags, gene sym-
bols, XML-like markup) EXCEPT the
pseudo-dialogue markers listed
above.
- Original wording, punctuation,
and order.  Do NOT fix grammar,
reflow text, or change casing.


OUTPUT: cleaned text ONLY (no
quotes, no notes).  If nothing
but commentary remains, output an
empty string.
```

Appendix B: wLAS for Base Models

Rank	lemma+UPOS	wLAS
1	result_NOUN	0.132
2	that_SCONJ	0.127
3	be_AUX	0.127
4	suggest_VERB	0.096
5	also_ADV	0.094
6	the_DET	0.079
7	than_ADP	0.060
8	significantly_ADV	0.048
9	10_NUM	0.034
10	show_VERB	0.034
11	conclusion_NOUN	0.031
12	mean_ADJ	0.031
13	may_AUX	0.030
14	high_ADJ	0.027
15	most_ADV	0.026
16	addition_NOUN	0.026
17	common_ADJ	0.023
18	._PUNCT	0.022
19	these_DET	0.022
20	a_DET	0.021

Table 4: Top-20 wLAS entries, aggregated over all base models. Higher values indicate greater overuse in model output relative to human continuations.

Appendix C: Model-Specific Results

Higher values indicate greater overuse in model output relative to human continuations.

Rank	lemma+UPOS	wLAS
1	be_AUX	0.142
2	result_NOUN	0.115
3	also_ADV	0.114
4	the_DET	0.107
5	that_SCONJ	0.085
6	10_NUM	0.055
7	suggest_VERB	0.053
8	than_ADP	0.051
9	significantly_ADV	0.045
10	show_VERB	0.039

Table 5: Top-10 wLAS for Falcon3 7B Base.

Rank	lemma+UPOS	wLAS
1	that_SCONJ	0.135
2	result_NOUN	0.127
3	be_AUX	0.119
4	suggest_VERB	0.109
5	also_ADV	0.078
6	the_DET	0.065
7	than_ADP	0.059
8	may_AUX	0.052
9	these_DET	0.048
10	addition_NOUN	0.040

Table 6: Top-10 wLAS for Gemma 3 4B Base.

Rank	lemma+UPOS	wLAS
1	that_SCONJ	0.148
2	result_NOUN	0.142
3	be_AUX	0.125
4	suggest_VERB	0.116
5	also_ADV	0.081
6	the_DET	0.067
7	than_ADP	0.052
8	may_AUX	0.041
9	these_DET	0.040
10	conclusion_NOUN	0.037

Table 7: Top-10 wLAS for Llama 3.1 8B Base.

Rank	lemma+UPOS	wLAS
1	result_NOUN	0.145
2	that_SCONJ	0.145
3	be_AUX	0.123
4	suggest_VERB	0.110
5	also_ADV	0.082
6	the_DET	0.071
7	than_ADP	0.070
8	significantly_ADV	0.055
9	conclusion_NOUN	0.046
10	10_NUM	0.040

Table 8: Top-10 wLAS for Mistral v0.3 7B Base.

Rank	lemma+UPOS	wLAS
1	be_AUX	0.141
2	also_ADV	0.139
3	result_NOUN	0.120
4	the_DET	0.100
5	that_SCONJ	0.086
6	suggest_VERB	0.066
7	than_ADP	0.056
8	significantly_ADV	0.052
9	a_DET	0.039
10	0_NUM	0.033

Table 9: Top-10 wLAS for OLMo 2 11B Base.

Rank	lemma+UPOS	wLAS
1	that_SCONJ	0.166
2	result_NOUN	0.146
3	suggest_VERB	0.122
4	be_AUX	0.115
5	than_ADP	0.072
6	also_ADV	0.068
7	the_DET	0.067
8	significantly_ADV	0.066
9	these_DET	0.051
10	may_AUX	0.048

Table 10: Top-10 wLAS entries for Yi 1.5 6B Base.

Bibliographical References

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

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad. arXiv preprint.

Gregory C. Allen. 2025. Deepseek: A deep dive. Congressional Testimony, CSIS.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, and et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Bryce Anderson, Riley Galpin, and Tom S Juzek. 2025. Model misalignment and language change: Traces of ai-associated language in unscripted spoken english. *arXiv preprint arXiv:2508.00238*.

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(6037).

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint.

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 Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NeurIPS*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Mitchell Cole. 2025. Deepseek's updated r1 ai model is more censored, test finds. *TechCrunch*.

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: Linguagem e Tecnologia*, 17:1–19.

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 *ACL 2025 Student Research Workshop*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *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 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, Florence, Italy. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

Elizabeth Gibney. 2025. China's cheap, open AI model deepseek thrills scientists. *Nature*, 638:13–14.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrew Gray. 2024. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*.

Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior*, 16(1):107–112.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Hugging Face Team. 2024. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

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.

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, Vienna, Austria. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Houji Jin, Negin Ashrafi, Armin Abdollahi, Wei Liu, Jian Wang, Ganyu Gui, Maryam Pishgar, and

Huanghao Feng. 2025. Llm encoder vs. decoder: Robust detection of chinese AI-generated text with LoRA. arXiv preprint arXiv:2509.00731.

Tom S Juzek and Zina B Ward. 2025. Word overuse and alignment in large language models: The influence of learning from human feedback. *arXiv preprint arXiv:2508.01930*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. 2024. On the reliability of watermarks for large language models. In *The Twelfth International Conference on Learning Representations*, ICLR 2024.

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. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*, pages 12–24.

Gabriela Kotz, Pedro Salcedo-Lagos, 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.

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

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

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*.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

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.

Ali Naseh, Donghyun Kang, Soroush Alizadeh, and Siamak Faridani. 2025. R1dacted: Investigating local censorship in deepseek's r1. *arXiv preprint arXiv:2505.12625*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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, Marseille, France. European Language Resources Association.

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.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

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, et al. 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*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434.

Billy Perrigo. 2023. Exclusive: Openai used kenyan workers on less than $2 per hour to make chatgpt less toxic. *TIME*.

Billy Perrigo. 2025. Is 'sweatshop data' really over? *TIME*.

Jacob Poushter, Maria Smerkovich, Moira Fagan, and Andrew Prozorovsky. 2025. Free expression seen as important globally, but not everyone thinks their country has press, speech and internet freedoms. Pew Research Center.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Christoph Riedl, Saiph Savage, and Josie Zvelebilova. 2024. AI's social forcefield: Reshaping distributed cognition in human–AI teams. *arXiv preprint arXiv:2407.17489*.

Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4):105–114.

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*.

Sigal Samuel and Jordan Crook. 2025. Here's how deepseek censorship actually works—and how to get around it. *WIRED*.

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*, 27:935–956.

M. Sharma et al. 2023. Towards understanding sycophancy in language models. *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.

Olivia Sidoti and Colleen McClain. 2025. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center.

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

Chris Stokel-Walker. 2025. We tried out deepseek. it worked well, until we asked it about tiananmen square and taiwan. *The Guardian*.

Team OLMo, Pete Walsh, Luca Soldaini, and et al. 2025. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Luka Terčon and Kaja Dobrovoljc. 2025. Linguistic characteristics of AI-generated text: A survey. arXiv preprint arXiv:2510.05136.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, 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.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc Le. 2025. Simple synthetic data reduces sycophancy in large language models. In *ICLR*.

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*.

Alex Young, Bei Chen, Chao Li, and et al. 2024. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*.

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, Monograph Supplement*, 9(2, Pt. 2):1–27.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing llm-generated and human-authored news text using formal syntactic theory. *arXiv preprint arXiv:2506.01407*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. 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. 2024. From lists to emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704*.