Why AI-Generated Text Is Harder to Understand Than You Think

Originally published 26 Oct 2026 at https://www.unite.ai/why-ai-generated-text-is-harder-to-understand-than-you-think/



AI models may be becoming harder to understand because their training increasingly rewards solving problems efficiently rather than explaining things clearly to humans.

The above, unedited sentence was written by ChatGPT in response to a request for an ELI5, single-sentence summary of a science paper; and it features at least a few of the language-mangling traits explained in that new study*. The new work concludes that this issue affects all frontier models, as well as near-frontier models, and some open-weights models that were also tested.

For instance, the early part of the segment *'their training increasingly rewards solving problems efficiently'* exemplifies the paper's stance that AI models sometimes *give non-living things human actions* (since 'training' is described in the above example as *actively rewarding something*).

This can impede quick parsing of the phrase, as we have to internally acknowledge the attribution as an illustrative method, rather than take it literally (which is much simpler).

The latter section *'increasingly rewards solving problems efficiently'* also squeezes several ideas together, forcing readers to work out what's being rewarded, what's being solved, and what's becoming more efficient.

The paper states†:

> *'[Readers] use animacy as a cue during parsing, and without additional context, a mismatch may elicit a P600 response†† associated with reanalysis.*
>
> *'However, rephrasing these sentences to avoid inanimate subjects would potentially require lengthening the sentence: for example, the word "bites" stands in for "causes an issue".*
>
> *'We speculate that during training, models may have converged to using such constructions as efficient, albeit non-standard, ways of expressing meaning which would otherwise take up more tokens.'*

Say What?

This is the core of the paper – that because modern AI models such as ChatGPT, Claude and Gemini are optimized for multiple types of problem-solving rather than specifically for communication, they tend to deliver language in the form of very compressed payloads that need some decoding.

This was a surprise to me, as I am constantly trying to persuade ChatGPT and its peers to be *less verbose*, and *more* compressed – a problem that is recognized and addressed in the literature.

Instead, the paper presents verbosity as a kind of zip-bomb that prevents easy understanding even in a sentence that is not very long. Therefore, combining the twin issues of excessive default reply length with the paper's observation about excessive compression, we can imagine that the characteristic pattern of frontier AI's default response style is quite distinct (and, presumably, this rather terse style is used in part as a determinant for AI-generated text in detectors, though the paper does not deal with this).

The authors also examine stacked noun compounds, exemplified in the problematic phrase *Squad Helps Dog Bite*†:

© Martin Anderson and/or commissioning entities

> *'In English, nouns freely modify other nouns in an adjectival way: in the phrase* coffee shop, coffee *modifies* shop, *despite usually appearing as a noun. This feature enables stacking multiple nouns into rich compound words which denote highly specific meanings.*
>
> *However, such stacks can introduce bracketing ambiguity, as in the headline Squad Helps Dog Bite Victim, and disrupt reading.'*

So if LLMs are verbose via compression, are they being *doubly* verbose by producing excessive *volumes* of compressed text in response to a query? That means that not only is there a great deal of text to read through in response to your query, but the text itself may have a lot of 'implicit' meaning that needs decoding.

One 2023 paper posited that LLMs can become unnecessarily verbose due to AI judges being used during training, since these often reward longer answers – even when humans preferred shorter ones, in the study's tests.

The authors of the new work attribute (see earlier quote) some of the compressive tendencies of frontier models to an architectural disposition to save tokens, which is a generous thought these days.

Let's take a look at the methods, data and tests run by the authors for the study.

Data and Methods

Unusual word combinations (measured through n-grams, or sequences of words) were investigated for their unpredictability, with rare combinations associated with greater reading difficulty.

Dependency length (the distance between grammatically-connected words) was also measured, since longer distances place greater demands on working memory – particularly when a subject and its verb are separated by several intervening words.

Object relative clauses (constructions such as *'the house the contractors built'*) were examined, particularly when optional words were omitted (such as *'that'* in *'the house that the contractors built'*).

The aforementioned *Stacked noun compounds* (three or more nouns joined together, as in *'horror movie soundtrack generator'*) were investigated as well, because the relationships between their component words can be ambiguous.

Denominal verbs (nouns used as verbs, such as *'fork it'* instead of *'put it on the fork')* were examined, alongside inanimate subjects paired with verbs normally associated with living things (as discussed earlier). Both constructions allow meanings to be compressed into fewer words, but additional interpretation may be required from readers.

The Contenders

The researchers collected AI-generated conversations from 14 models; Kimi K2.5; GLM 5; Claude Sonnet 5, Opus 5, Opus 5.5, and Fable 5; GPT-5.6 Luna, Terra and Sol; Olmo Base, Instruct and Think; and Llama Base and Instruct.

The tests comprised ten starting prompts for general conversation, and five for coding, with each task requiring a model to clone a real public GitHub repository, find bugs, and fix them in a sandbox.

Conversations were limited to 12 exchanges, while coding sessions were limited to 11, followed by a fixed closing message. In a separate test, each model was asked to write seven essays based on assignments from the PERSUADE 2.0 dataset.

For comparison, human-written texts were drawn from sources matched as closely as possible to the three types of AI output: OpenSubtitles (a collection of film and television subtitles unrelated to the open source subs platform) was used for general conversation; StackExchange (the online network of question-and-answer forums) for coding discussions; and student essays from the aforementioned PERSUADE 2.0 dataset for the writing tasks. Wikipedia was also used as an additional comparison source.

Where suitable human data was available, the same linguistic features were measured across human and AI-generated text, with these reference collections also used to assess the frequency of word combinations.

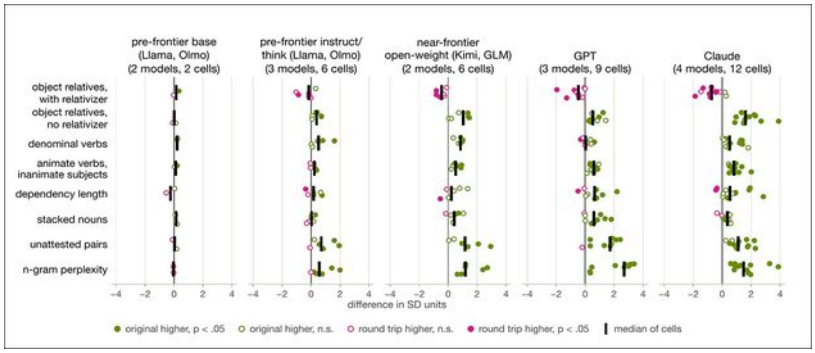
Round-trip translation (English to Japanese and back, using the NLLB-200 model) was used to distinguish stylistic preferences from linguistic necessity. Meaning was largely preserved while sentence structure was changed, allowing unnecessarily complex wording to be identified.

Results

Experiment 1

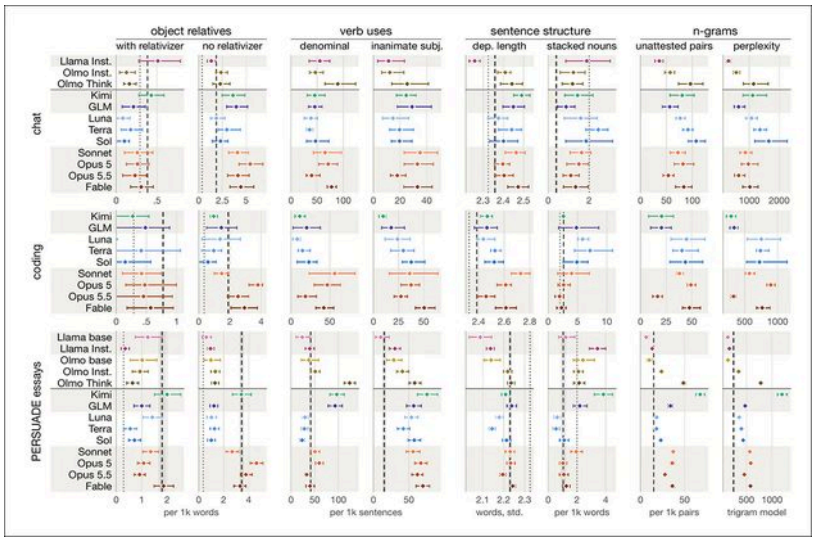
As shown below, several difficult constructions were found more frequently in the original AI-generated text, particularly among frontier models. This was interpreted as evidence of a preference for unnecessarily complex phrasing:

© Martin Anderson and/or commissioning entities



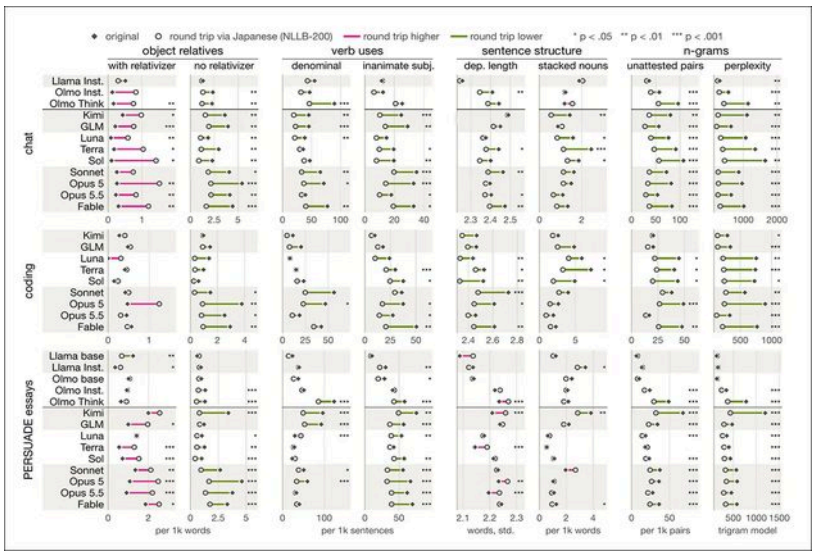
Comparison of original AI-generated text with versions translated into Japanese and back into English. Green dots show language patterns used more often in the patterns used more often after translation. The differences were generally greater for GPT and Claude models. Source

The six linguistic features were compared across 14 AI models in three settings (chat, coding and essay writing). Twelve models were tested in chat, nine in coding and all 14 in essay writing:



Comparison of six linguistic features across AI models in chat, coding and essay writing. Each diamond represents a model's pooled result; horizontal bars show 9 vertical lines indicate human-written baselines from OpenSubtitles (chat), StackExchange (coding) and PERSUADE 2.0 student essays (essay writing). Dotted line comparisons. Higher values generally indicate more frequent use of the measured construction, or greater linguistic complexity. Please refer to the source paper fo

The original AI-generated text was also compared with its retranslated versions, allowing differences in sentence structure to be measured:



Comparison of original AI-generated text (solid diamonds) with versions translated into Japanese and back into English using NLLB-200 (hollow circles). Pink lines frequently after translation; green lines indicate features used more frequently in the originals. Asterisks indicate statistical significance (one, two or three asterisks statistical evidence). Most constructions associated with reading difficulty were reduced by translation, particularly among frontier models. Please refer to the sour

Frontier and near-frontier models were found to over-use unusual word combinations, denominal verbs, inanimate subjects and object relative clauses with optional words omitted (see graph above). These tendencies were generally weaker in Llama and Olmo Base, but were more pronounced in Olmo Think.

© Martin Anderson and/or commissioning entities

Smaller differences were observed in essay-writing than in chat or coding, which was attributed to the single-prompt nature of essays, with greater linguistic compression potentially encouraged by extended back-and-forth conversations.

The authors note:

> *'The PERSUADE essays show smaller changes under round-trip translation on several categories (n-gram measures, dependency lengths, stacked nouns, denominal verbs) relative to chat and [coding].*
>
> *'We hypothesize that this is because they involve a single prompt, unlike the many iterated turns of chat and coding. We speculate that certain patterns may emerge in long back-and-forth conversations (which may demand more information compression).'*

The least change after translation was found in Llama and Olmo Base and Instruct, which also came closest to student essays in their word combinations and verb usage. In contrast, Olmo Think was found to behave more like the frontier models. Claude Opus 5.5 was observed to use fewer difficult constructions than Opus 5 in 19 out of 21 comparisons, with non-overlapping 95% confidence intervals in eight of those.

Experiment 2

The second experiment was designed to test whether an underlying pressure to produce shorter and more information-dense responses might cause models to prefer linguistic constructions harder for humans to understand.

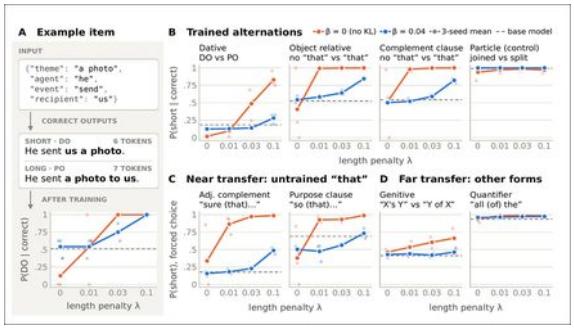
Here, Qwen2.5-1.5B-Instruct was fine-tuned using Group Relative Policy Optimization (GRPO), and the model tasked with converting structured, JSON-formatted descriptions of events into English sentences.

The two aspects of training adjusted were how much the model was allowed to change its usual choice of words (*β*); and how strongly it was penalized for producing longer responses (*λ*). Each event could be expressed in two grammatically-valid ways, with a known difference in token count.

For example, *'He sent us a photo'* required six tokens, compared with seven for *'He sent a photo to us'*. A separate test was also used in which two ways of saying the same thing required the same number of tokens, to check whether differences were caused by sentence length.

The effect was tested on previously unseen sentences using four constructions encountered during training; two untrained constructions involving the optional word 'that'; and unrelated constructions involving possessives, quantifiers, passives and purpose clauses.

The tests were designed to establish whether a preference for shorter wording would be learned and carried over into constructions never encountered during training:



The effect of penalizing longer responses during training. An example of shorter and longer phrasing is shown in A. Preferences for shorter constructions are measured in B, with tests of previously untrained constructions shown in C and D. Orange indicates training without a restriction on changes to the original model; blue indicates training with that restriction. Dashed lines show the original model's preferences. Stronger penalties generally produced greater preference for shorter wording, even in constructions absent from training. Please refer to the source paper for better resolution.

As the penalty for longer responses was increased, shorter wording was chosen more often (illustrated directly above). Without the KL penalty (a restriction on how far the model could depart from its original word predictions), the shift was rapid; with it, the same effect was observed more gradually. Without either penalty, preferences varied widely between training runs, but even a *small* length penalty produced a consistent preference for shorter alternatives.

This preference was also observed in previously unseen sentence constructions involving the optional word *'that'*. Transfer to more distant grammatical constructions was much weaker, with an effect found only for possessives when the KL penalty was absent.

The authors conclude:

> *'We take these results as a proof of concept that a length penalty used during reinforcement learning with GRPO can directly affect grammatical features.*
>
> *'We note that the model was not directly instructed to use compressed or dense language, and was not directly given information about the availability of syntactic alternations.*
>
> *'Rather, the penalty resulted in an emergent preference for shorter members of syntactic alternations.'*

Conclusion

I have been interested in the anchors – both inadvertent and intentional – that characterize AI-generated text for some time. Therefore this new offering sheds some light on the apparent conflict between LLMs' noted verbosity and their now equally-noted tendency to *compress* information in ways that seem 'easy to read', simply because not many words are used – but this is clearly the wrong reading of that cue.

© Martin Anderson and/or commissioning entities

In LLMs, brevity in text output may signal complexity rather than essential distillation, as if the AI were handing us a puzzle instead of the direct concepts behind the words. All, apparently, so that a single model can code *and* write.

* *Titled 'Why is AI-generated language so hard to understand?', from two authors across New York University and The University of Texas.*

† *My conversion of the authors' inline citations to hyperlinks.*

†† https://mikeydoes.com/glossary/p600/

First published Monday, October 26, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai