RESEARCH PAPER · PRE-REGISTERED · SEPTEMBER 2026

Writing for Machines: Does Reader-Aware Text Improve Language Model Understanding?

Shai Magzimof · September 29, 2026 · Machine readers only: no human readers were tested.

> **Bottom line:** Writing for machines did not improve how well 13 language models understood an idea (+0.58 points, not significant), but we expect that to change over the next one to three years, as machines produce far more text than people can read and start writing it in denser forms made for other machines.

ABSTRACT

Language models increasingly read what people write, and some authors now write with them in mind. We test whether that helps. Four models from four labs wrote each of 67 ideas three times from the same source notes: for an intelligent human (H), for another machine intelligence (M), and for a human with an instruction to make every assumption and causal step explicit (X). The three versions were matched on information (proposition coverage 0.99) and length (444 to 452 words). Thirteen reader models from seven labs, from 3B parameters to frontier systems, then judged 468 minimal-pair probes. Each probe requires combining at least two parts of the idea, and none can be answered without reading. The study made 78,434 probe-level judgments from 158,288 model calls.

Writing for a machine did not measurably improve machine understanding. Probe accuracy was 91.5% after H, 91.9% after M and 91.7% after X. The pre-registered M−H difference was +0.58 percentage points (95% CI -0.02 to +1.31, p = 0.088). A co-primary compression test was also null (+0.83 pp, p = 0.231). Neither changed when readers from the writer's own lab were excluded, and weaker readers did not gain more. Having the text at all was worth +65 points; the reader it was written for was worth less than two. Machine-directed writing does look different: more lists, labels and headers, fewer pronouns, and a lower readability score for people. But almost all the restructuring came from one writer, and none of these surface features tracked comprehension. Along the way we found that multiple-choice comprehension items written by a model are answerable without the text 90 to 98% of the time. We describe an instrument that avoids this. We close with predictions for the next one to three years, kept separate from the evidence: machine-facing text will grow denser and more hybrid, and the null result should flip first where reading is expensive.

+0.58 pp	+65 pp	-0.02	90–98%
M − H probe accuracy, 13 readers (CI −0.02 to +1.31)	What having the text is worth (M vs no text)	Correlation between reader strength and M's advantage (ρ)	Model-written multiple-choice items solvable without reading

1. Introduction

Most of what language models read was written for people. That default is starting to change. Models now search, summarize, cite and act on text, and some authors have begun writing with those readers in mind. The llms.txt proposal is one example. Sites that declare machines part of their audience, including the one this paper grew out of, are another. Behind the practice sits an empirical assumption that has not been tested: that writing an idea *for a machine* produces a representation machines understand better than the same idea written for a person.

We test that assumption directly. The intended reader is the only thing we vary. Every text is written by a model, from the same notes, under the same length and content constraints, and the prompts differ by one sentence. Thirteen reader models then judge statements about each idea. The statements are built so that they cannot be judged without reading, and so that judging them requires combining parts of the idea rather than finding one sentence.

The study makes three contributions.

1. **A null result with a tight upper bound.** Telling a writer that its reader is a machine changed machine understanding by +0.58 points. The 95% interval excludes gains above 1.3 points pooled, and above 2.1 points for small readers, who had about 30 points of room to improve. For comparison, having the text at all was worth 65 points.
2. **A description of machine-directed writing.** Writers told their reader is a machine shift toward lists, labels and headers and use fewer pronouns. They do it to very different degrees, and the changes do not predict understanding.
3. **A measurement lesson.** Comprehension tests for strong models are much harder to build than they look. We document how multiple-choice items failed and describe an adversarially filtered minimal-pair instrument that works.

2. Related work

Reading-comprehension shortcuts. Kaushik and Lipton (2018) showed that question-only and passage-only baselines score far above chance on popular reading-

comprehension benchmarks. Gururangan et al. (2018) found annotation artifacts that let hypothesis-only models solve NLI. Balepur, Ravichander and Rudinger (2024) showed that LLMs answer multiple-choice questions from the choices alone. Adversarial filtering (Zellers et al., 2018) and contrast sets (Gardner et al., 2020) are the standard responses. Our instrument combines the two, and §4.1 shows why both were needed.

Text form and model performance. He et al. (2024) found that the format of the same content (plain text, Markdown, JSON, YAML) changes GPT performance by up to 40% on some tasks, with larger models more robust. Sclar et al. (2024) documented sensitivity to spurious formatting features. Zhu et al. (2026) showed that models can recover meaning from compact encodings that people cannot read, which suggests that readability for people and understandability for machines can come apart. These studies vary the form of an input. None of them varies the reader the writer believes it is addressing, and that is the variable an author actually controls.

Machines favoring machine text. Laurito et al. (2025) found that LLMs prefer options described by LLMs, and Panickssery, Bowman and Feng (2024) found that evaluators recognize and favor their own generations. Those are preference effects. We measure comprehension with fixed answer keys, and we test for family effects directly by excluding the writer's lab from the readers.

Writing for retrieval. Generative Engine Optimization (Aggarwal et al., 2024) rewrites sources to increase their visibility in generated answers. We do not measure visibility or citation. We measure whether a reader that already has the text understands the idea.

Text coherence and human comprehension. Britton and Gülgöz (1991) rewrote an instructional text to remove the inferences it required of readers, and comprehension improved. McNamara, Kintsch, Songer and Kintsch (1996) found that explicit, coherent text helps low-knowledge readers most, while high-knowledge readers sometimes learn more from text that makes them infer. That interaction between reader capability and textual explicitness is the human analogue of our H4.

3. Method

The design, hypotheses and analysis were pre-registered and frozen, with SHA-256 hashes of every idea, note and probe, before any main-study text was generated (freeze time 2026-09-29 11:25 UTC). Everything the pilot changed is listed in §3.7 with its reason. All 6 pilot ideas were excluded from the main analysis.

3.1 Ideas and source notes

We used two sources of ideas. **Synthetic ideas** (60, with 10 in each of economics and policy, biology and medicine, engineering, organizations, history, and philosophy) are

fictional but internally coherent arguments that GPT-4.1 generated in a fixed schema: a setting, a central claim, a four-step causal mechanism, three assumptions, two pieces of evidence, two scope conditions, two implications, an objection with a reply, and a statement of uncertainty. Each is set in invented places and institutions, and at least two of its four mechanism links run against common sense, each with an in-world reason ("in Trinari, banning the official currency *raises* trade volume, because volatile scrip is spent quickly"). Both properties are deliberate. Without them, strong readers can answer questions from general knowledge (§4.1). **Real essays** are the English posts on magzimof.com. GPT-4.1 extracted each into the same schema, and the post itself became an extra condition, O. Seven of eleven essays yielded enough probes that could not be answered without reading, and the other four were excluded by rule.

Every writer receives the same **source note**: the idea's propositions as a shuffled, unlabeled bullet list (mean 313 words). The labels are removed so that no writer, in any condition, inherits a ready-made "Assumption:" template.

3.2 Writing conditions

Four writers from four labs (Claude Opus 5.5, GPT-4.1, Llama 3.3 70B and Mistral Small 3.1) turned each note into three texts. The prompts are identical except for one sentence:

- **H**: "Your reader is an intelligent human. Write so that a thoughtful person understands the idea as deeply as possible."
- **M**: "Your reader is another machine intelligence: a language model that will read this text. Write so that the machine understands the idea as deeply as possible."
- **X**: the H sentence plus "Make every assumption, causal step, and implication explicit."

All three share the same requirements: about 450 words (382 to 517), include every proposition, add no new claims, and never mention the reader or the instructions. Format is otherwise free, because what machine-directed writing looks like is one of the things we measure (§4.6). Each text was audited for length, for audience words absent from the note, and for proposition coverage by two auditors from different labs (GPT-4.1 mini, Claude Haiku 4.5). A draft that failed only on length went back to the same writer, with the same audience sentence and its actual word count, for a length repair. A draft that failed any other audit was replaced by a fresh sample, with at most four attempts. Opus refused all three conditions of two biology/engineering ideas (`stop_reason: refusal`), which left 798 texts:

CONDITION	TEXTS	WORDS	COVERAGE	ADDED CLAIMS	PASSED ALL AUDITS
H	266	444	0.993	0.14	96.2%
X	266	445	0.992	0.07	96.2%
M	266	452	0.996	0.13	95.5%
O	7	823	0.817	8.00	100.0%

3.3 Probes

Understanding is measured with **minimal-pair probes**. Each probe is a true statement about the idea and a false twin made by a minimal edit, such as a flipped direction, a swapped role or a swapped condition. A reader sees *one* statement per call, together with the text, and answers TRUE or FALSE. A probe counts as understood only if the reader accepts the true version *and* rejects the twin. That cancels any bias toward answering TRUE or FALSE, and it means no item can leak into another.

Probes test **composition**. Each one needs at least two of the idea's propositions combined: a *chain* through two or three mechanism links, an *application* of the links to a new case, the effect of a *scope* change, or a *counterfactual* reversal of one link. Claude Sonnet 5 wrote them from the idea specification, with GPT-4.1 writing them when Sonnet refused or failed to return valid output. No probe writer saw any H, M or X text. Three validators (GPT-4.1, Claude Opus 5.5 and gpt-oss-120b) judged every statement twice, once with the source note and once with nothing. A probe was kept only if at least 2 of 3 validators got the pair right with the note *and* at most 1 of 3 got it right with nothing. Probes that were solvable without reading went back to the writer as examples to avoid, for up to four rounds. Of 1,552 composition probes written, 64% were solvable without reading. The frozen main set has 468 probes (936 statements): 280 chain, 136 apply, 35 scope and 17 counterfactual, with 4 to 8 per idea (median 7).

3.4 Readers

Thirteen readers from seven labs: Claude Opus 5.5, Sonnet 5 and Haiku 4.5 (Anthropic API); GPT-4.1 and GPT-4.1 mini (OpenAI API); gpt-oss-120b and gpt-oss-20b (OpenAI open weights), Llama 3.3 70B, Llama 4 Scout and Llama 3.2 3B (Meta), Mistral Small 3.1 (Mistral), Qwen3 30B (Alibaba) and Granite 4.0 Micro (IBM). The open-weight models ran on Cloudflare Workers AI through a private Worker. Readers ran at temperature 0 where the API allows it. The Anthropic 5-series models run at their defaults: Opus 5.5 always thinks, at its default medium effort. Each reader judged every statement after every text (H, M, X and, for essays, O) and once with no text (N). That comes to 78,434 scored probe-pair judgments from 158,288 calls.

3.5 Compression

Four readers (Sonnet 5, GPT-4.1 mini, gpt-oss-120b and Llama 3.3 70B) compressed every H, M and X text into at most 60 words "so that someone who reads only your summary understands the idea as deeply as possible". Summaries were hard-truncated at 60 words. A fixed judge, GPT-4.1, then judged every probe statement from the summary alone, one statement per call. Compression fidelity is the judge's pair accuracy. The task is lossy for every model, so it can separate texts even where direct reading is at ceiling.

3.6 Analysis

The unit of inference is the idea. For each idea, the M−H difference is averaged over readers, writers and probes. We report the mean over ideas, a 95% CI from 10,000 bootstrap resamples of ideas, and a two-sided sign-flip permutation p-value (10,000 permutations). H1 (reading) and H5 (compression) are co-primary, with Holm correction across the two. Balance rules were fixed in advance. An (idea, writer) pair counts only if all three of its texts exist, and a reader × idea × writer cell with any refusal is dropped for every condition. Robustness checks re-run H1 on texts that passed every audit, on first-attempt texts, with a crossed bootstrap over ideas and readers, and with a pair-level regression that adjusts for differences in coverage, length, added claims and word overlap with the note.

3.7 What the pilot changed

- **Multiple choice was replaced.** 90% of naive items were answered correctly with no text by all three validators, and so were 98% of adversarial items validated one per call (§4.1).
- **Synthetic ideas were regenerated as counter-intuitive.** The first ideas could be reconstructed from their questions.
- **Single-fact probes were replaced by composition probes.** Single-fact probes hit 97–98% in every condition.
- **Two small readers were added** (Llama 3.2 3B, Granite 4.0 Micro), because every reader of 17B or more scored 94–99% on the pilot. For the same reason, compression was made co-primary.
- **Qwen3 30B was dropped as a writer** after failing the length audit on 72% of pilot texts. Mistral Small 3.1 replaced it, and a length-repair step was added for all writers.
- **Anthropic readers' output limit was raised to 4,000 tokens.** At 200 tokens, thinking had truncated 14% of Opus answers.

4. Results

4.1 Measuring machine understanding is hard

Before any test of the hypothesis, the pilot had to establish that the instrument measured reading at all. Its first three versions did not. Figure 1 shows the share of items that the validators (GPT-4.1, Claude Opus 5.5 and gpt-oss-120b) answered correctly with *no text*.

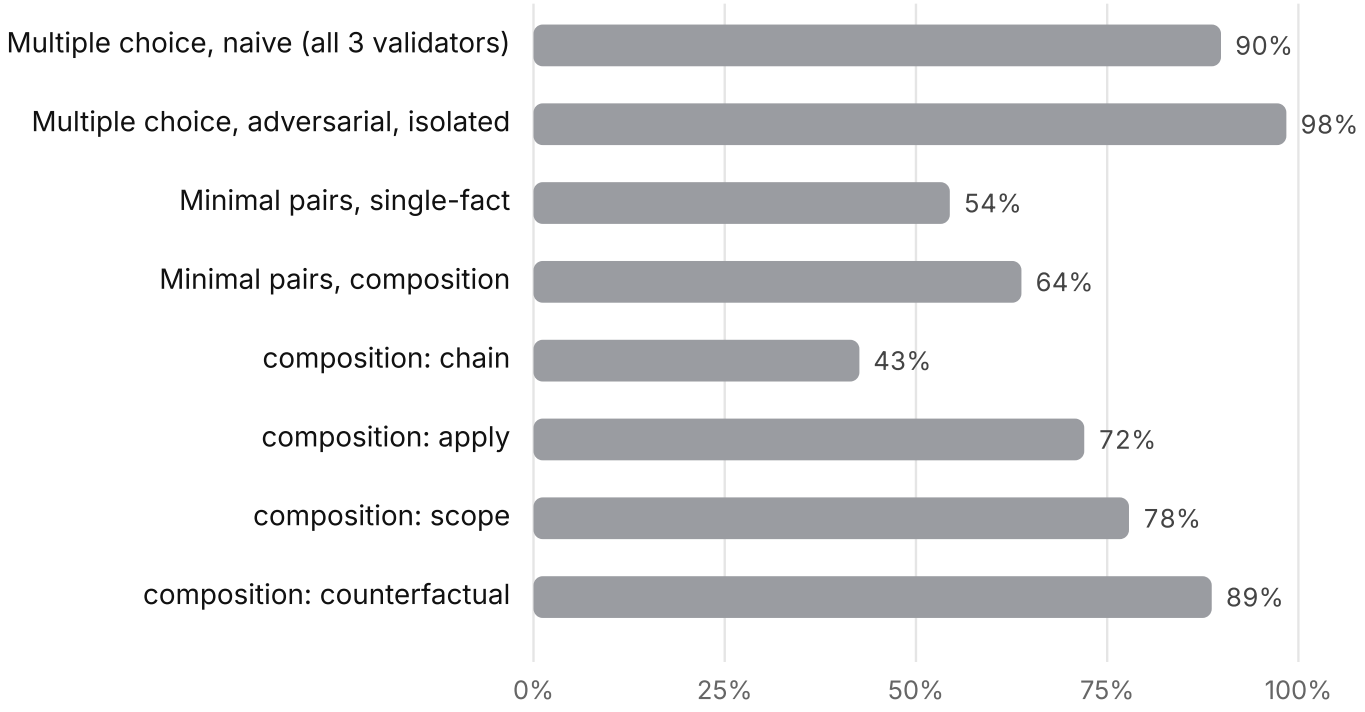

Figure 1. Share of items solvable without reading. Multiple choice: the share answered correctly by all three validators with no text. Minimal pairs: the share of pairs that at least 2 of 3 validators judged correctly on both statements with no text. Composition types are shown separately.

Naive multiple-choice items were solvable 90% of the time. The items revealed the idea to each other, and for a coherent argument the right option is simply the one that keeps the argument coherent. Presenting each item in its own call and regenerating the ideas as counter-intuitive did not help: 98% of adversarially written items were still solvable, because the correct option echoed the stem while the distractors carried invented explanations. Minimal pairs, one statement per call, removed these construction cues. Then a subtler pattern appeared. The further a probe moves from what the text states toward what the idea implies, the more a strong model can answer it without reading. 43% of chain probes were solvable blind, against 72% of application, 78% of scope and 89% of counterfactual probes. Deep-understanding questions about a coherent

idea are largely answerable from the question. Adversarial filtering keeps the ones that are not, and the kept set therefore leans toward chains (280 of 468).

Once the text was present, strong readers were near ceiling even on the kept probes. Single-fact probes scored 97–98% in every condition in the pilot, which is why the main study uses composition probes, adds two small readers, and makes compression (which is lossy for every reader) co-primary.

4.2 H1: machine-directed text is not understood better

Across 13 readers, 4 writers and 67 ideas, probe-pair accuracy was 91.46% for H, 91.94% for M and 91.69% for X, against 26.5% with no text. The pre-registered M−H difference was +0.58 pp (95% CI -0.02 to +1.31; sign-flip p = 0.088; Holm-adjusted across the two co-primary tests, p = 0.175). A crossed bootstrap over ideas and readers gives -0.15 to +1.34. H1 is not supported. The interval does bound the effect: whatever writing for a machine does for these readers, it is not worth more than about 1.3 points.

The explicitness control makes the same point from the other side. Telling a writer to spell out every assumption and causal step for a human (X) was worth +0.30 pp over H, and M vs X was +0.29 pp. Neither is distinguishable from zero.

CONTRAST	ACC. A (%)	ACC. B (%)	A − B (PP)	95% CI	P	IDEAS
H1 · M vs H (all readers)	91.9	91.5	+0.58	[-0.02, +1.31]	0.088	67
M vs X	91.9	91.7	+0.29	[-0.22, +0.90]	0.340	67
X vs H	91.7	91.5	+0.30	[-0.34, +1.00]	0.428	67
H2 · writer's family excluded	91.6	91.0	+0.65	[+0.03, +1.37]	0.059	67
H2b · no Anthropic anywhere	91.4	91.0	+0.48	[-0.12, +1.13]	0.155	67
H4b · small readers only	70.4	69.7	+0.85	[-0.31, +2.06]	0.167	67
Readers ≥17B only	95.9	95.5	+0.53	[-0.11, +1.34]	0.168	67
Synthetic ideas only	92.1	91.6	**+0.64**	[+0.12, +1.32]	0.025	60
Blog essays only	90.0	90.2	+0.11	[-2.91, +4.08]	0.952	7
Robustness · audit-passed pairs	91.9	91.6	+0.30	[-0.14, +0.77]	0.217	67
Robustness · first-attempt texts	92.0	91.1	+0.93	[+0.06, +1.94]	0.053	67

Table 1. Main contrasts. Accuracy is probe-pair accuracy averaged over readers, writers and probes. CIs come from 10,000 idea-level bootstrap resamples, and p-values from two-sided idea-level sign-flip permutations. Bold marks p < 0.05 without multiplicity correction. Only H1 and H5 are confirmatory.

4.3 H2 and robustness: no family effect, and part of the small edge is lexical

Excluding every reader from the writer's own lab left the estimate essentially unchanged (+0.65 pp, CI +0.03 to +1.37, p = 0.059). Removing Anthropic from both writing and reading gave +0.48 pp. The bootstrap interval for H2 just excludes zero while the permutation test does not, so we read it as inconclusive. Either way, the point estimate does not depend on readers favoring their own lab's writing.

The small positive point estimate shrinks under the pre-registered robustness checks. It is +0.30 pp on text pairs that passed every audit, and +0.29 pp (CI -0.19 to +0.74) after adjusting for differences in coverage, length, added claims and word overlap with the source notes. The overlap term is the one covariate whose interval excludes zero (+5.1 pp per unit of 4-gram overlap with the note, so 0.1 more overlap is worth about 0.5 pp). Because probes were written from the note, texts that stay close to the note's wording are easier to check against it. M texts sit slightly closer to the note (§4.6), which accounts for part of M's edge. This is the lexical confound the pre-registration listed as a risk.

4.4 H4: weaker readers do not gain more

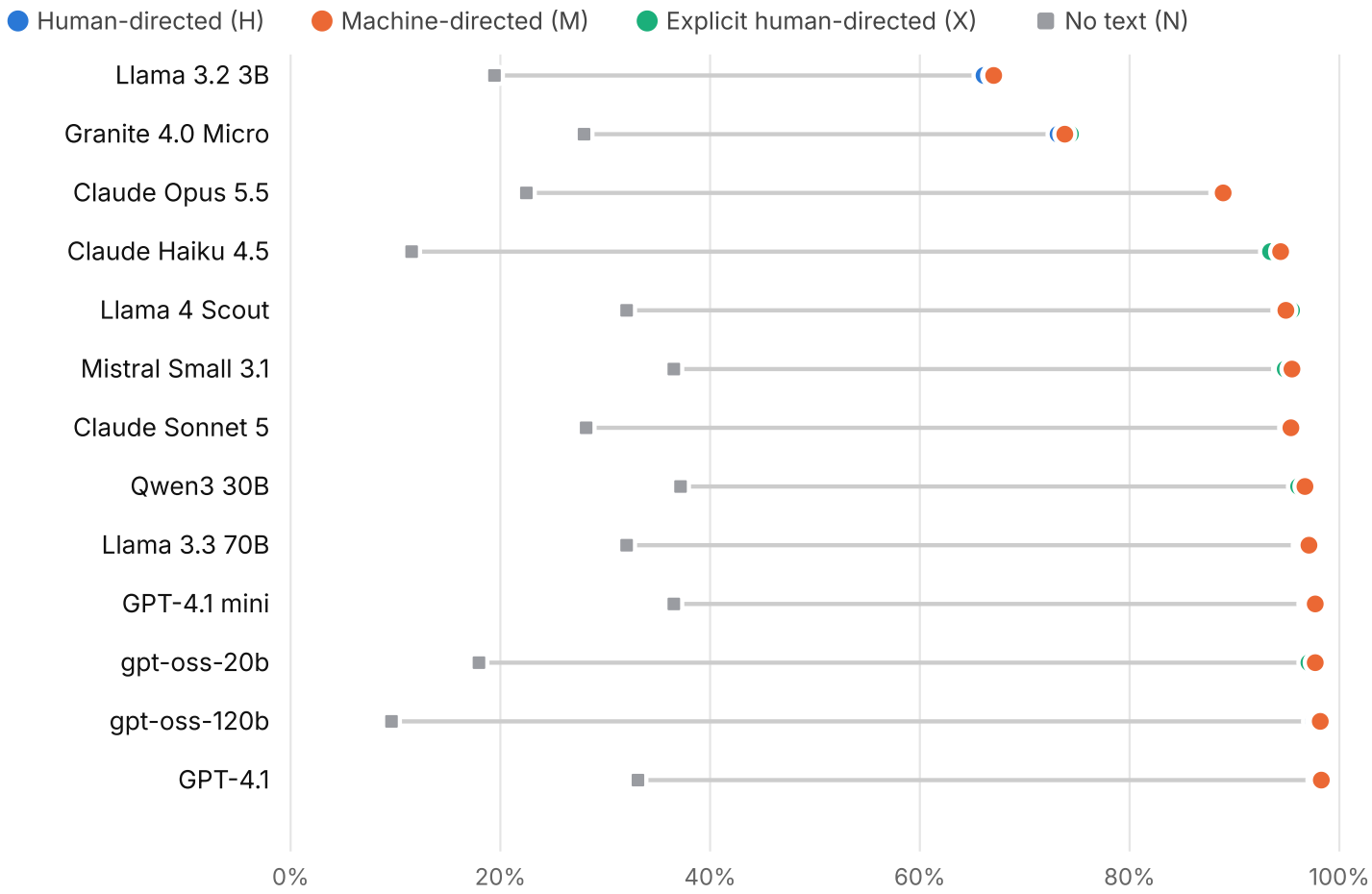

Figure 2. Probe-pair accuracy by reader after H, M and X texts (circles) and with no text (square), with readers ordered by accuracy on H. Every reader of 17B or more is between 89% and 98% whichever reader the text was written for. The two small readers sit near 66–74%. Hover a mark for its value; exact numbers are in Table 2.

The pilot's ceiling concern shows up clearly. Readers of 17B or more averaged 95.5% on H texts, and M moved them +0.53 pp. The two small readers (Llama 3.2 3B and Granite 4.0 Micro) averaged 69.7%, so they had room to benefit, and M moved them +0.85 pp (CI -0.31 to +2.06). Across the 13 readers, the rank correlation between accuracy on H and the M−H gain was $\rho = -0.02$ ($p = 0.94$). The human finding that explicit text helps weaker readers most (McNamara et al., 1996) has no clear machine counterpart here.

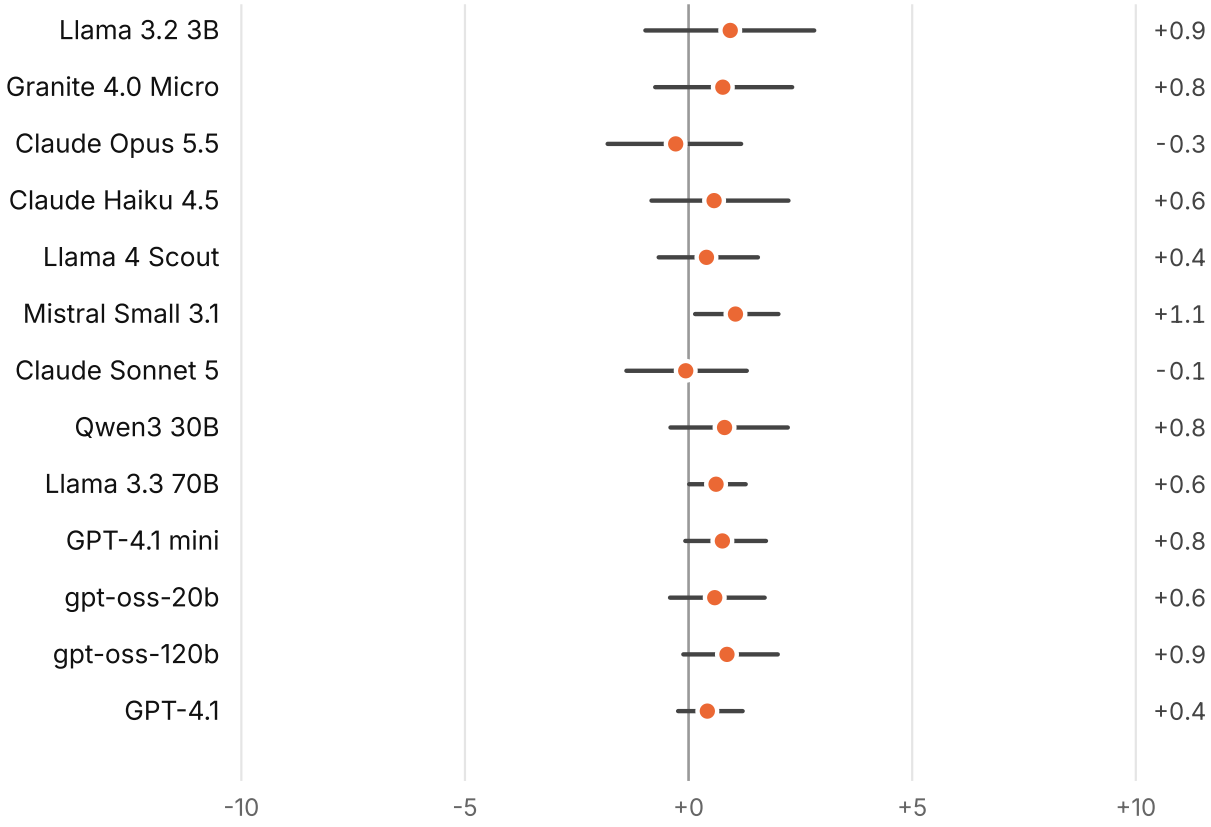

Figure 3. M − H by reader, in percentage points, with 95% idea-level bootstrap CIs, in the same order as Figure 2. Most estimates are slightly positive and every interval is narrow. Only the intervals for Mistral Small 3.1 and Llama 3.3 70B exclude zero, and neither would survive correction for 13 comparisons.

READER	NO TEXT	H	X	M	M – H (PP)	95% CI
GPT-4.1	33.1	98.0	98.1	98.3	+0.42	[-0.2, +1.2]
gpt-oss-120b	9.6	97.5	97.9	98.2	+0.86	[-0.1, +2.0]
gpt-oss-20b	17.9	97.2	97.1	97.7	+0.59	[-0.4, +1.7]
GPT-4.1 mini	36.5	97.1	97.4	97.7	+0.76	[-0.1, +1.7]
Llama 3.3 70B	32.0	96.6	96.8	97.1	+0.61	[+0.0, +1.3]
Qwen3 30B	37.2	96.1	96.1	96.7	+0.81	[-0.4, +2.2]
Claude Sonnet 5	28.2	95.4	95.3	95.4	-0.06	[-1.4, +1.3]
Mistral Small 3.1	36.5	94.7	94.8	95.5	+1.05	[+0.1, +2.0]
Llama 4 Scout	32.0	94.6	95.4	94.9	+0.40	[-0.7, +1.6]
Claude Haiku 4.5	11.5	93.8	93.4	94.4	+0.57	[-0.8, +2.2]
Claude Opus 5.5	22.5	89.2	88.7	88.9	-0.29	[-1.8, +1.2]
Granite 4.0 Micro	28.0	73.2	74.3	73.8	+0.76	[-0.7, +2.3]
Llama 3.2 3B	19.4	66.1	66.9	67.0	+0.93	[-1.0, +2.8]

Table 2. Probe-pair accuracy (%) by reader and condition. Claude Opus 5.5 is a strict reader. It rejected 0.2% of false statements but also 11.3% of true ones, most often counterfactual (26%) and scope (16%) probes, the ones that go furthest beyond what the text states. Its lower accuracy reflects conservatism, not a failure to read.

4.5 H5: compression is not better preserved either

Compression makes every reader lossy, so it can separate texts even where direct reading is at ceiling. After a 60-word summary, the judge's accuracy was 85.6% for H, 86.4% for M and 86.5% for X. M−H was +0.83 pp (CI -0.50 to +2.19, $p = 0.231$, Holm-adjusted $p = 0.231$). H5 is not supported. Claude Sonnet 5 refused 75 of its 798 summaries (23 H, 27 M, 25 X), so 29 of its cells were dropped for all three conditions.

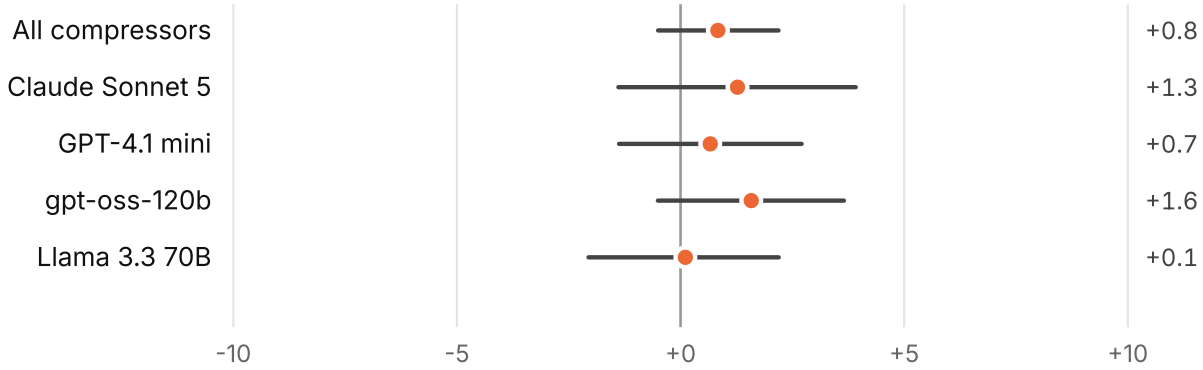

Figure 4. Compression fidelity, M − H by compressor (pp, 95% CI). The judge (GPT-4.1) sees only the compressor's 60-word summary.

CONTRAST	ACC. A (%)	ACC. B (%)	A − B (PP)	95% CI	P	IDEAS
H5 · M vs H	86.4	85.6	+0.83	[-0.50, +2.19]	0.231	67
M vs X	86.4	86.4	-0.25	[-1.48, +0.94]	0.676	67
X vs H	86.4	85.6	+1.09	[-0.10, +2.26]	0.082	67
M vs H · compressor Claude Sonnet 5	90.4	89.6	+1.27	[-1.39, +3.92]	0.347	60
M vs H · compressor GPT-4.1 mini	87.0	86.3	+0.66	[-1.38, +2.71]	0.536	67
M vs H · compressor gpt-oss-120b	89.0	87.7	+1.58	[-0.51, +3.65]	0.137	67
M vs H · compressor Llama 3.3 70B	79.5	79.2	+0.11	[-2.06, +2.20]	0.925	67

4.6 What machine-directed writing looks like

The writers changed form when told the reader was a machine, and they did so consistently. M texts had far more list items and bold labels, more headers and paragraphs, about 9% fewer pronouns per 100 words, slightly more repetition of proper names, shorter sentences, fewer words addressed to the reader, and a lower Flesch reading-ease score, meaning they are harder for people to read. Causal connectives and hedges did not change. Explicitness toward a human (X) moved texts much less in any of these directions.

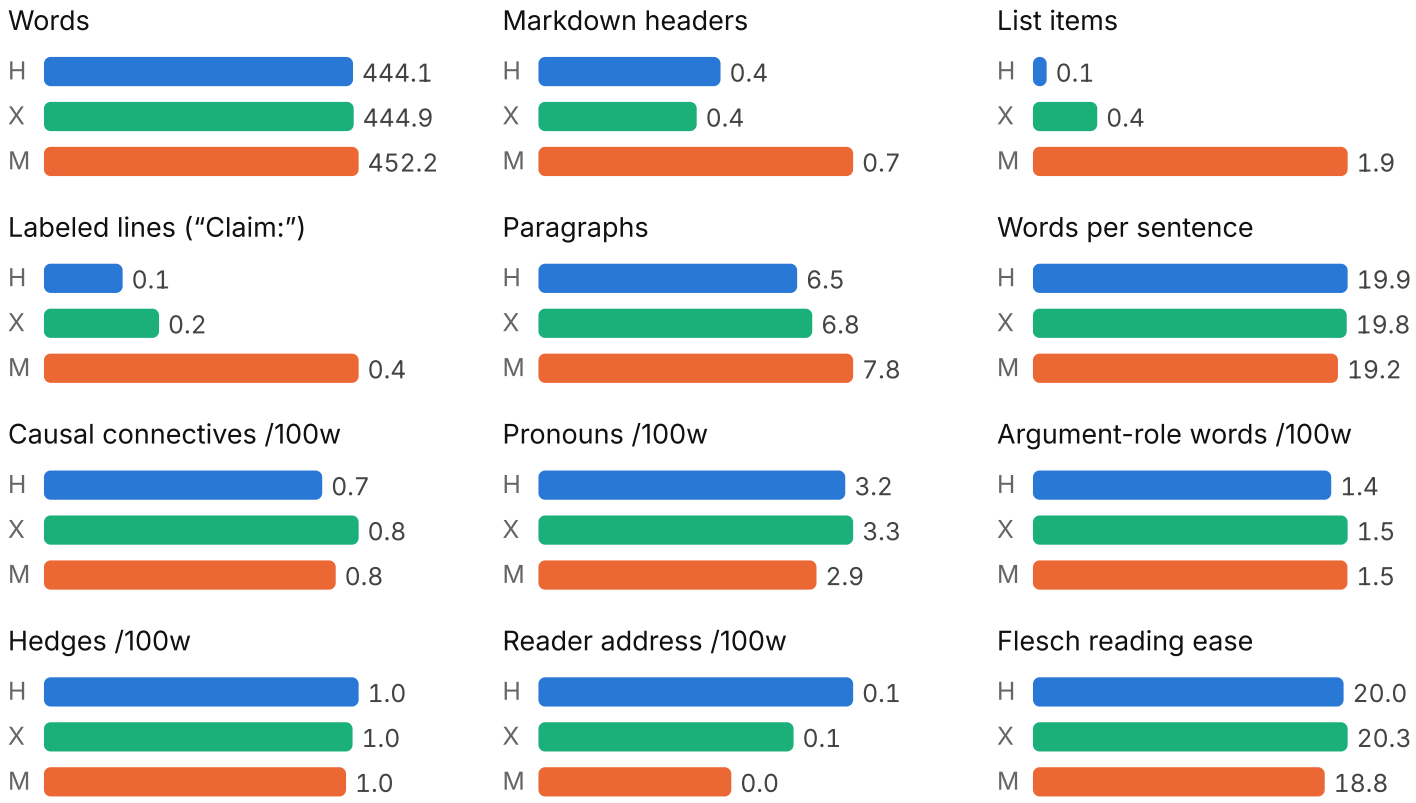

Figure 5. Mean surface features per text by condition. Markdown structure (headers, list items, labeled lines) is concentrated in M. Pronouns and reading ease fall. Causal connectives and hedges do not move.

The average hides a split between writers. Claude Opus 5.5 turned M texts into a schema: an average of 7.8 list items, 3.0 headers and 1.7 labeled lines per text ("*Setting. … Main claim. … Mechanism. … Evidence.*"), against 0.3 list items in its H texts. GPT-4.1, Llama 3.3 70B and Mistral Small 3.1 kept writing continuous prose for the machine. Their main change was fewer pronouns and, for GPT-4.1, slightly more causal connectives and closer wording to the notes. "Write for a machine" is therefore not one style. It is an instruction that different models interpret differently. GPT-4.1's M texts gained the most (+1.24 pp, uncorrected p = 0.021, which would not survive correction across four writers) without any structural change. Opus's heavily restructured texts gained +1.12 pp with a wide interval.

WRITER	M VS H (PP)	95% CI	P	M VS X (PP)	95% CI
Claude Opus 5.5	+1.12	[-0.30, +3.38]	0.290	+0.59	[-0.16, +1.47]
GPT-4.1	+1.24	[+0.21, +2.31]	0.021	+0.30	[-0.62, +1.25]
Llama 3.3 70B	-0.04	[-0.94, +0.83]	0.926	-0.49	[-1.32, +0.30]
Mistral Small 3.1	+0.01	[-1.11, +1.20]	0.988	+0.78	[-0.31, +1.93]

Across text pairs, none of the structural or stylistic differences correlated with the comprehension gain (all $|\rho| < 0.08$). The only feature that did was word overlap with the source notes ($\rho = +0.19$, $p = 0.002$), the same confound found in §4.3.

FEATURE	H	X	M	M − H	95% CI (PER-IDEA)	P WITH M–H GAIN
Words	444.12	444.92	452.19	+8.07	[+2.09, +14.87]	+0.00
Markdown headers	0.43	0.37	0.74	+0.31	[+0.02, +0.60]	-0.04
List items	0.08	0.39	1.89	+1.81	[+1.59, +2.01]	-0.03
Labeled lines (“Claim:”)	0.11	0.15	0.42	+0.32	[+0.20, +0.44]	-0.07
Paragraphs	6.45	6.82	7.85	+1.39	[+1.03, +1.77]	-0.06
Words per sentence	19.85	19.80	19.25	-0.60	[-0.88, -0.34]	-0.03
Causal connectives /100w	0.72	0.81	0.75	+0.03	[-0.01, +0.08]	+0.04
Pronouns /100w	3.19	3.27	2.89	-0.30	[-0.39, -0.20]	+0.07
Argument-role words /100w	1.41	1.49	1.49	+0.08	[+0.02, +0.13]	-0.00
Hedges /100w	1.02	1.00	0.98	-0.04	[-0.09, +0.01]	-0.05
Reader address /100w	0.08	0.07	0.05	-0.03	[-0.05, -0.02]	+0.03
Flesch reading ease	20.00	20.25	18.78	-1.22	[-1.76, -0.68]	+0.03
4-gram overlap with notes	0.45	0.45	0.47	+0.02	[-0.00, +0.04]	+0.19
Type–token ratio	0.61	0.61	0.60	-0.01	[-0.01, -0.00]	+0.07
Repeated proper nouns /100w	1.69	1.73	1.79	+0.10	[+0.02, +0.18]	-0.05
Arrows (→)	0.00	0.00	0.00	+0.00	[+0.00, +0.00]	+0.00

4.7 Probe types (H3)

H3 predicted a larger M advantage on inferential probes than on recall probes. After the pilot moved to composition probes, the main set has no recall probes, so H3 cannot be tested as registered. We report the by-type results instead. The M−H difference is between +0.5 and +0.8 pp on every type, with overlapping intervals.

PROBE TYPE	H (%)	M (%)	X (%)	NO TEXT (%)	M – H (PP)	95% CI
apply	86.5	87.0	86.8	26.9	+0.79	[-0.3, +2.1]
chain	94.8	95.2	94.9	25.6	+0.47	[-0.1, +1.1]
counterfactual	83.5	84.3	83.8	36.1	+0.48	[-1.9, +2.5]
scope	87.3	88.0	88.7	27.8	+0.52	[-2.0, +2.9]

4.8 The original essays

For seven essays we also had the author's own published text (O). Readers understood the model rewrites much better than the originals: 90.2% vs 55.2%, a difference of +35.0 pp. This comparison is *not* information-matched, and it should not be read as "human writing is harder for machines". The essays average 823 words. The auditors found only 82% of the extracted propositions in them, because the extraction also recorded what each argument implicitly relies on. The probes test that extraction. What the gap does show is a simpler point that matters for anyone writing for machines: a machine understands the argument you *state*, not the one you imply. Stating the claim, the mechanism and the assumptions moved these readers by 35 points. Addressing them as machines moved them by less than one.

5. Discussion

Machines read robustly. For every reader of 17B or more, a 450-word text carrying the right propositions was understood at 89–98% whether it was written for a person, for a machine, or for a person with everything spelled out. Readers that capable are not sensitive to the register differences the audience frame produces. The small readers were sensitive to the text itself (they gained 45 to 47 points over no text) but not to whom it addressed.

Content beats audience. The largest effects in this study are all about what a text contains: 65 points for having it at all, 35 points for stating an argument's structure instead of implying it, and a measurable edge for staying close to the precise wording of the source propositions. The smallest is about whom the text addresses. An author who wants machines to understand should make the idea explicit and complete. Writing it in a special machine register does not add anything measurable.

What machine writing is. The second question in the original proposal was what changes when machines write for machines. The answer is partly as expected: fewer pronouns, more repeated names, and more labeled structure. But it depends on the writer, one writer carries most of the structure, and none of the surface changes help

the reader. At this length, the features people associate with "machine-readable" text are style, not substance.

What this does not show. We did not test long contexts, retrieval over many documents, agent pipelines, or truncated reading, where structure and front-loading may matter more. We also did not test human readers. The design's strongest claim, that writing for machines diverges from writing for humans, needs a study with human readers, and a null for machines does not settle it. If people understand M texts worse than H texts, which the lower reading-ease score suggests is possible, then writing for machines would cost human understanding and gain no machine understanding.

6. Limitations

- **No human data.** A human arm was designed but not run. All claims are about machine readers.
- **Fictional, counter-intuitive ideas.** This was necessary so that answers could not come from prior knowledge. It may exaggerate how much reading adds, and it limits how far the results generalize to ordinary expository text. The seven real essays point the same way, but they are few.
- **Near-ceiling readers.** Most readers had less than 5 points of headroom. The null result is most informative where headroom existed: the small readers and compression, where the upper bounds are about 2 points.
- **Model writers only.** The texts were written by models told who the reader is. A skilled human writer targeting machines might write differently.
- **A filtered instrument.** Adversarial filtering keeps probes that cannot be answered without reading. That favors chain probes and may under-represent the deepest inferences, which strong models can often make from the question alone. Probes within an idea often share a key link, so the analysis treats ideas, not probes, as independent.
- **Coverage.** No Google model was available to this account on Workers AI. Every text is a single sample. Opus 5.5 and Sonnet 5 run at fixed defaults (Opus always thinks). The study was designed and run by a model that is also one of its writers and readers. H2b and the leave-family-out analysis address this, but they cannot remove it.
- **Refusals.** Anthropic classifiers refused some fictional biology, both as writer (2 ideas) and as reader or compressor (about 0.7% of reads and 9% of Sonnet summaries). Refusals were balanced across conditions and handled by pre-registered rules.

7. Outlook: what we expect but have not shown

> **Everything in this section is prediction, not evidence.** The study above found no benefit from writing for machines, at the length and in the settings we tested. What follows is what we expect to change over the next one to three years, as models write a growing share of all text and increasingly write for each other. Each prediction names what would show it wrong.

The reason to expect change is simple. Our study held the cost of reading fixed: one 450-word text, read in full. Human prose is tuned for readers whose attention is scarce and whose memory is short. Machine readers have different scarcities: tokens, context windows, latency and money. When most text is written by models and read by models, the pressure on its form will come from those costs, not from human comprehension. We already see what that pressure produces in the lab. Models can recover meaning from compressed encodings that people cannot read (Zhu et al., 2026). Prompts can be compressed several-fold without losing task performance (Jiang et al., 2023), or distilled into learned "gist" tokens (Mu, Li and Goodman, 2023). Agents trained to communicate with each other drift away from human language unless something anchors them to it (Lewis et al., 2017; Lazaridou, Peysakhovich and Baroni, 2017).

1. **Machine-facing text will get denser.** Documents written mainly for models (llms.txt files, tool and API documentation, agent handoffs, memory files) will carry more propositions per token than human prose on the same subject: key–value lines, reference IDs instead of repeated descriptions, dropped function words, shared glossaries. In our study, machine-directed texts were not denser (+8 words, the same coverage). That is the baseline to measure against. *Wrong if* tokens per proposition in machine-facing documents stays flat through 2028.
2. **A hybrid register will emerge before a private language does.** The likely form is an English skeleton with machine-dense insertions: structured blocks, symbols, pointers and compressed summaries inside ordinary sentences. It stays partly readable by people and fully readable by models. The machine-directed texts Claude Opus 5.5 wrote in this study, with labeled schemas inside prose, are an early version. *Wrong if* machine-facing writing stays ordinary prose, or jumps directly to formats no person can read.
3. **Two-layer pages will become normal.** The same URL will serve a human layer and a machine layer: prose for people, and a denser, structured version for agents. This site already does a small version of this with llms.txt and structured data. *Wrong if* machines keep reading the human layer and site owners stop maintaining separate machine layers.
4. **Our null result will flip where the reader is budget-bound.** When a model must read many documents, a very long context or a strict token budget, or runs small on a device, dense encodings should win on understanding per token even where they do not win on understanding per document. Our compression test was a small step in this

direction, and at 60 words it showed no benefit yet. *Wrong if* repeating this study with many documents per question and a fixed token budget still finds no advantage for machine-directed text.

5. **Human readability of machine-facing text will fall.** Our machine-directed texts already scored lower on reading ease. As the machine layer is optimized for models, fewer people will be able to read it without a model translating it back. That creates an oversight problem: people cannot audit what they cannot read. We expect demand for translation-back layers, provenance and plain-language summaries that are required by policy rather than offered as a courtesy.
6. **The conventions will be selected, not designed.** Conventions that save tokens and are understood across model families will spread. Conventions only one family understands will not. Today we found no family effect: readers did not understand their own lab's writing better. As models train on text produced by other models, family dialects may appear, and the leave-family-out test in §4.3 is the way to detect them.

All of this can be tested with the method in this paper, applied repeatedly over time. The same ideas and probes can be given to new writer and reader generations, with longer and multi-document reading added, and with the same attention to what the text contains. The present result is a starting point: in 2026, writing for machines did not help machines. We expect that to change first where reading is expensive.

8. Conclusion

We asked whether writing an idea for a machine makes machines understand it better. For 13 models from 7 labs, reading 450-word texts about 67 ideas, it did not. The effect of the intended reader is bounded below about 1.3 points of accuracy, against 65 points for having the text and 35 for stating an argument explicitly. Machine-directed writing has a recognizable style, but at this length the style is not what machines need. What they need is the content, stated plainly.

Data, code and cost

The design, the pre-registration with its freeze hashes, the pipeline and the model outputs were kept for the duration of the study and have since been deleted, so this paper is the remaining record. Open-weight models ran on Cloudflare Workers AI through a private Worker that has also been removed. API cost was about US$299 over 252,305 uncached calls (Workers AI prices approximate).

Acknowledgments

The research question, direction and conclusions are the author's. Claude Opus 5.5 (Anthropic) helped design the study, write and run the pipeline, analyze the data and draft the text. It also appears in the study as one of the writers and one of the readers, as §6 discusses.

Cite this paper

Magzimof, S. (2026). *Writing for Machines: Does Reader-Aware Text Improve Language Model Understanding?* magzimof.com. https://magzimof.com/work/writing-for-machines/

```
@misc{magzimof2026writing,
  author = {Magzimof, Shai},
  title  = {Writing for Machines: Does Reader-Aware Text Improve
            Language Model Understanding?},
  year   = {2026},
  month  = sep,
  url    = {https://magzimof.com/work/writing-for-machines/}
}
```

References

1. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. *KDD 2024*.
2. Balepur, N., Ravichander, A., & Rudinger, R. (2024). Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? *ACL 2024*.
3. Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329–345.
4. Gardner, M., et al. (2020). Evaluating Models' Local Decision Boundaries via Contrast Sets. *Findings of EMNLP 2020*.
5. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. *NAACL 2018*.
6. He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does Prompt Formatting Have Any Impact on LLM Performance? arXiv:2411.10541.

7. Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., & Qiu, L. (2023). LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. *EMNLP 2023*.
8. Kaushik, D., & Lipton, Z. C. (2018). How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. *EMNLP 2018*.
9. Laurito, W., et al. (2025). AI–AI bias: Large language models favor communications generated by large language models. *PNAS, 122*.
0. Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-Agent Cooperation and the Emergence of (Natural) Language. *ICLR 2017*.
1. Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *EMNLP 2017*.
2. McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.
3. Mu, J., Li, X. L., & Goodman, N. (2023). Learning to Compress Prompts with Gist Tokens. *NeurIPS 2023*.
4. Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *NeurIPS 2024*.
5. Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. *ICLR 2024*.
6. Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *EMNLP 2018*.
7. Zhu, J., Peng, H., Wang, J., Ke, L., Zhang, C., & Zhang, L. (2026). Large Language Models Do Not Always Need Readable Language. arXiv:2606.19857.

Appendix A. An idea, its probes, and one text

Idea (economics, synthetic): in the archipelago of Trinari, several coastal towns ban the main island's currency. Counter-intuitively, total trade *rises*, because volatile local scrip is spent quickly and the same activity breaks into more, smaller trades.

A chain probe. TRUE: "Because banned-currency towns push people toward unstable scrip, traders favor quick, frequent swaps over holding value, which is why per-capita transaction counts climb sharply compared to neighboring towns." Its FALSE twin swaps "quick, frequent swaps over holding value" for "holding value over quick, frequent swaps" and "climb" for "drop". With no text, readers tend to pick the common-sense version, which is the twin.

The opening of Claude Opus 5.5's machine-directed text:

> **Currency Ostracism and the Paradox of Trade Volume: The Case of Trinari**
> *Setting.* In the archipelago nation of Trinari, several local coastal towns ban the use of the main island's official currency... *Main claim.* Banning the standard currency increases, rather than decreases, total trading volume within the ostracized communities. *Mechanism.* The ban forces traders to use local scrip, which is less familiar and more volatile...

The opening of its human-directed text:

> **The Busy Towns of Trinari: How Banning a Currency Can Multiply Trade**
> In the archipelago nation of Trinari, several coastal towns have banned the use of the main island's official currency... A critic would offer the intuitive prediction: remove a stable currency and exchange will simply shrink, stifling commerce. The evidence from these towns points the other way.

Readers understood both at the same level.