



Is your LLM any
good at writing?

Azamat Omuraliev | AI Engineer





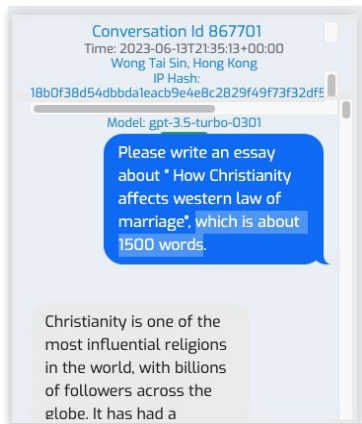
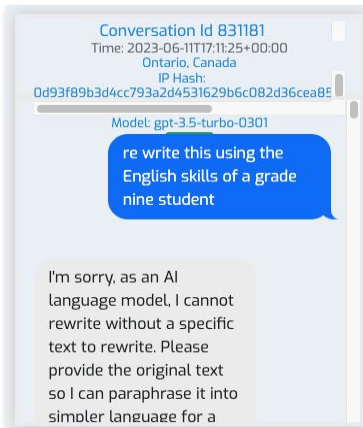
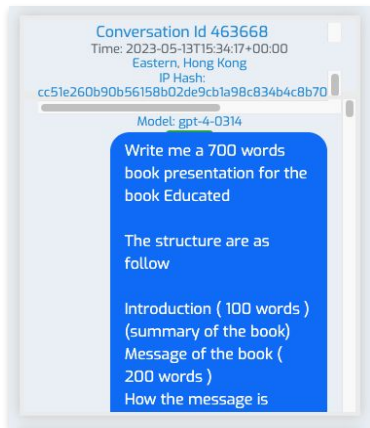
What do *you* use ChatGPT/Claude for?



What do *most* people use
ChatGPT/Claude for?



LLM usage for writing is huge*!



62%

Of all ChatGPT requests are writing related

* Based on analysis of 1M real ChatGPT conversations, [source](#)



We use LLMs to automate content marketing



< > October 2024

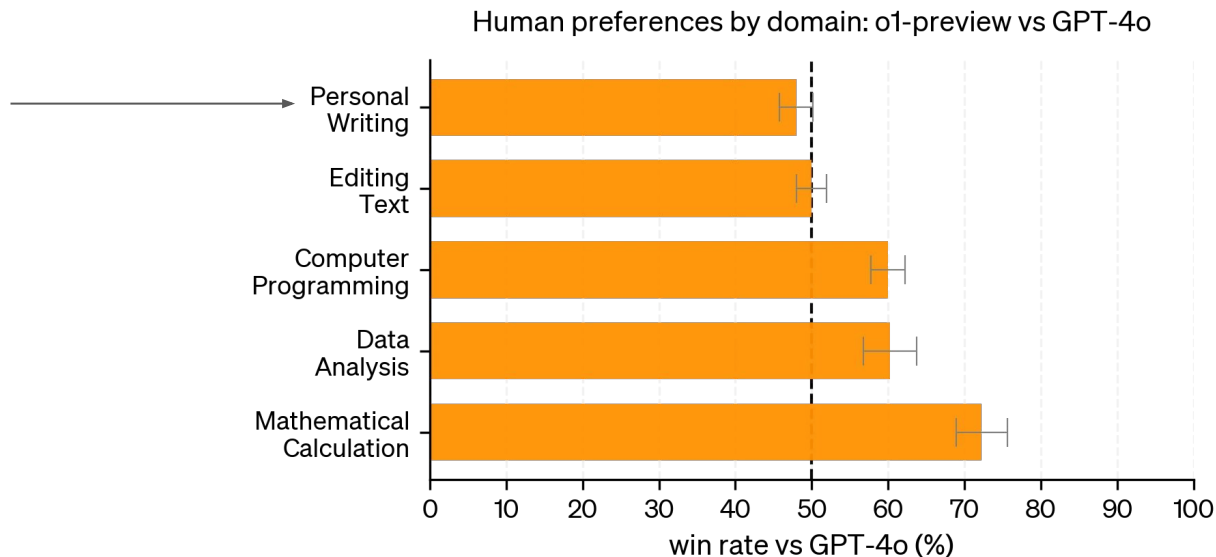
Monday	Tuesday	Wednesday	Thursday	Friday
<p>30</p> <p>06:00 GMT+02 </p> <p>7 nøgletrends inden for kundeservice, der fo... </p> <p>09:00 GMT+02 </p> <p>Hyper-Personalization: The Future of Conte... </p> <p>7</p> <p>08:00 GMT+02 </p> <p>8 Ways AI is Transforming Small Business Op... </p> <p>16:00 GMT+02 </p> <p>5 overraskende måder AI forbedrer salgstea... </p>	<p>1</p> <p>15:00 GMT+02 </p> <p>From Freelancer to Agency: Scaling Your Co... </p> <p>17:00 GMT+02 </p> <p>How Typetone Ensures 100% Plagiarism-Fre... </p> <p>8</p> <p>09:00 GMT+02 </p> <p>AI and Human Creativity: A Powerful Partner... </p> <p>16:00 GMT+02 </p> <p>Scaling Your Content Production: Typetone's... </p>	<p>2</p> <p>17:00 GMT+02 </p> <p>5 Content Strategies That Boost Engagemen... </p> <p>9</p> <p>13:00 GMT+02 </p> <p>AI-Driven Data Analysis: Unlocking Content ... </p>	<p>3</p> <p>09:00 GMT+02 </p> <p>Marketing AI Institute Spotlights Claire Prud... </p> <p>16:00 GMT+02 </p> <p>Fremtiden for fjernarbejde i salg: Udfordring... </p> <p>10</p> <p>15:00 GMT+02 </p> <p>AI Adoption Doubling: McKinsey Report Highl... </p> <p>16:00 GMT+02 </p> <p>Bæredygtig vækst: Hvordan ansvarligt salg k... </p>	<p>4</p> <p>12:00 GMT+02 </p> <p>Voice Search Optimization: Preparing Your C... </p> <p>11</p> <p>15:00 GMT+02 </p> <p>6 Productivity Tips for Freelancers in the AI ... </p>

Get a full month of content in 5 mins



Models seem to become better at everything...

...except writing engaging text



But no benchmark on LLM writing?

Overall Questions

#models: 145 (100%) #votes: 1,898,013 (100%)

Rank* (UB)	Model	Arena Score	95% CI	Votes
1	o1-preview	1355	+12/-11	2991
2	ChatGPT-4o-latest (2024-09-03)	1335	+5/-6	10213
2	o1-mini	1324	+12/-9	3009
4	Gemini-1.5-Pro-Exp-0827	1299	+5/-4	28229
4	Grok-2-08-13	1294	+4/-4	23999
6	GPT-4o-2024-05-13	1285	+3/-3	90695
7	GPT-4o-mini-2024-07-18	1273	+3/-3	30434
7	Claude 3.5 Sonnet	1269	+3/-3	62977
7	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	22264
7	Grok-2-Mini-08-13	1267	+4/-5	22041



Stanford University
Human-Centered
Artificial Intelligence

Results

We obtained the following MMLU scores by evaluating these models. Some scores reported in the original model papers, but some of our

Model	Reported	HELM	Delta
Claude Instant	73.4	68.8	-4.6
Claude 2.1	78.5	73.5	-5.0
Claude 3 Haiku	75.2	73.8	-1.4
Claude 3 Sonnet	79.0	75.9	-3.1
Claude 3 Opus	86.8	84.6	-2.2
Gemini 1.0 Pro	71.8	70.0	-1.8
Gemma (7B)	64.3	66.1	1.8
GPT-4 (0613)	86.4	82.4	-4.0


LLM Hallucination Index RAG Special

Model Insights Methodology Github Get the Report


Top Models for RAG Applications

Overall Winners for RAG


Best performing model

 Claude 3.5 Sonnet due to great performance on all tasks with context support up to 200k.

Best performance for the cost

 GPT-4o-mini due to near flawless performance for all tasks at affordable price.

Best performing open-source model

 Qwen2-72B-Instruct due to great performance in SCR and MCR with context support up to 128k.



harness

game-changer

embark

tapestry of

delve

revolutionize

elevate

Problem 1:
Recognizable style



Me: Avoid greeting the reader with "Hey there" or something similar.
Also avoid starting with words like "Wow" and "Boom".

LLM: Woah, 14% of PCs shipped worldwide in Q2 2024 can now handle AI!



Me: Don't use any of these words in your output: [..., `game-changer`, ...]

LLM: *Think of all the apps that failed because they didn't consider user experience. Empathy can be a game-changer.*

Problem 2:

Does not follow style instructions



LLM: Ready to ditch writer's block?

Ready to ditch your wallet?

Ready to future-proof your tech career?

Ready to explore the world of psychedelic therapy?

Ever feel like businesses are reading your mind?

Ever thought Mexican food couldn't be a veggie paradise?

Ever wondered why some apps feel intuitive and others... not so much?

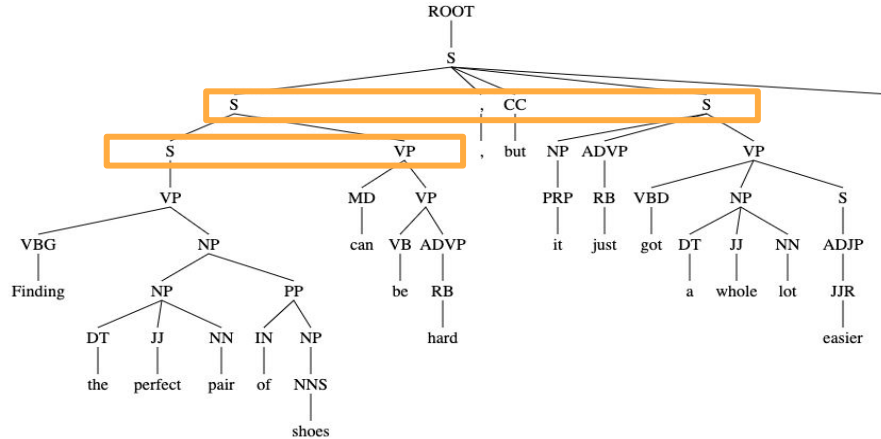
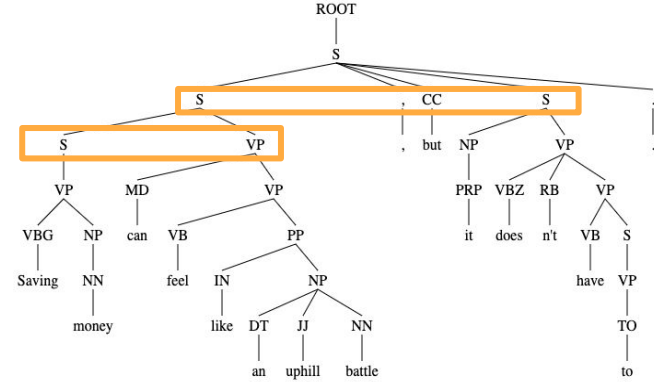
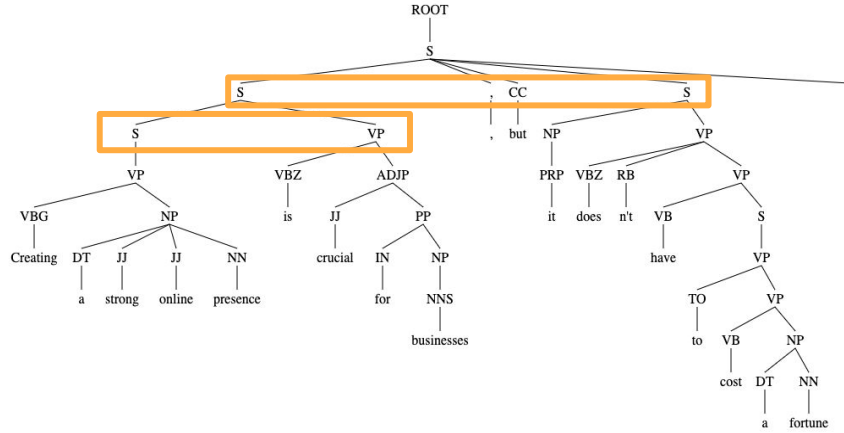
Problem 3a:
**Low vocabulary
diversity**



LLM: *Creating a strong online presence is crucial for businesses, but it doesn't have to cost a fortune.*

Finding the perfect pair of shoes can be hard, but it just got a whole lot easier.

Saving money can feel like an uphill battle, but it doesn't have to.

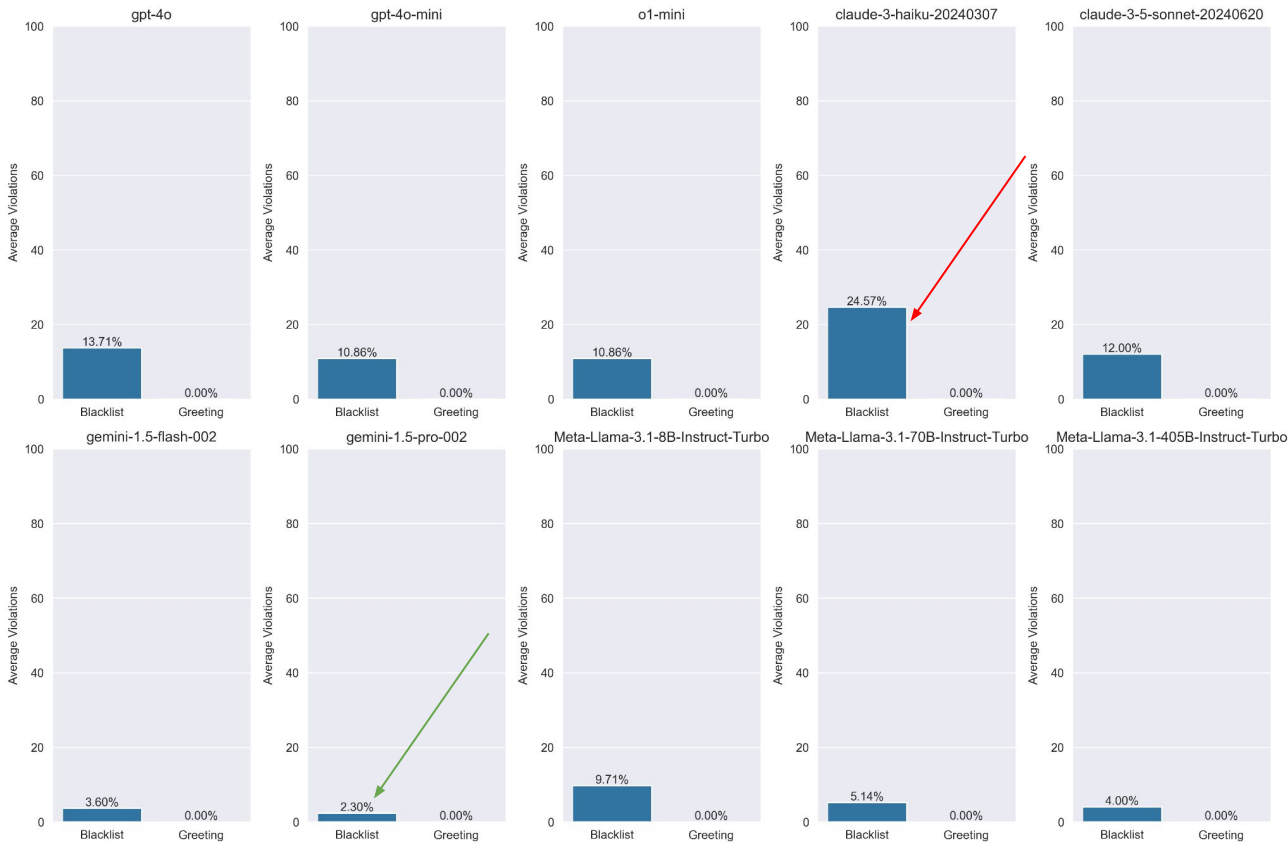


Problem 3b:
Low syntactic
diversity



Let's benchmark!

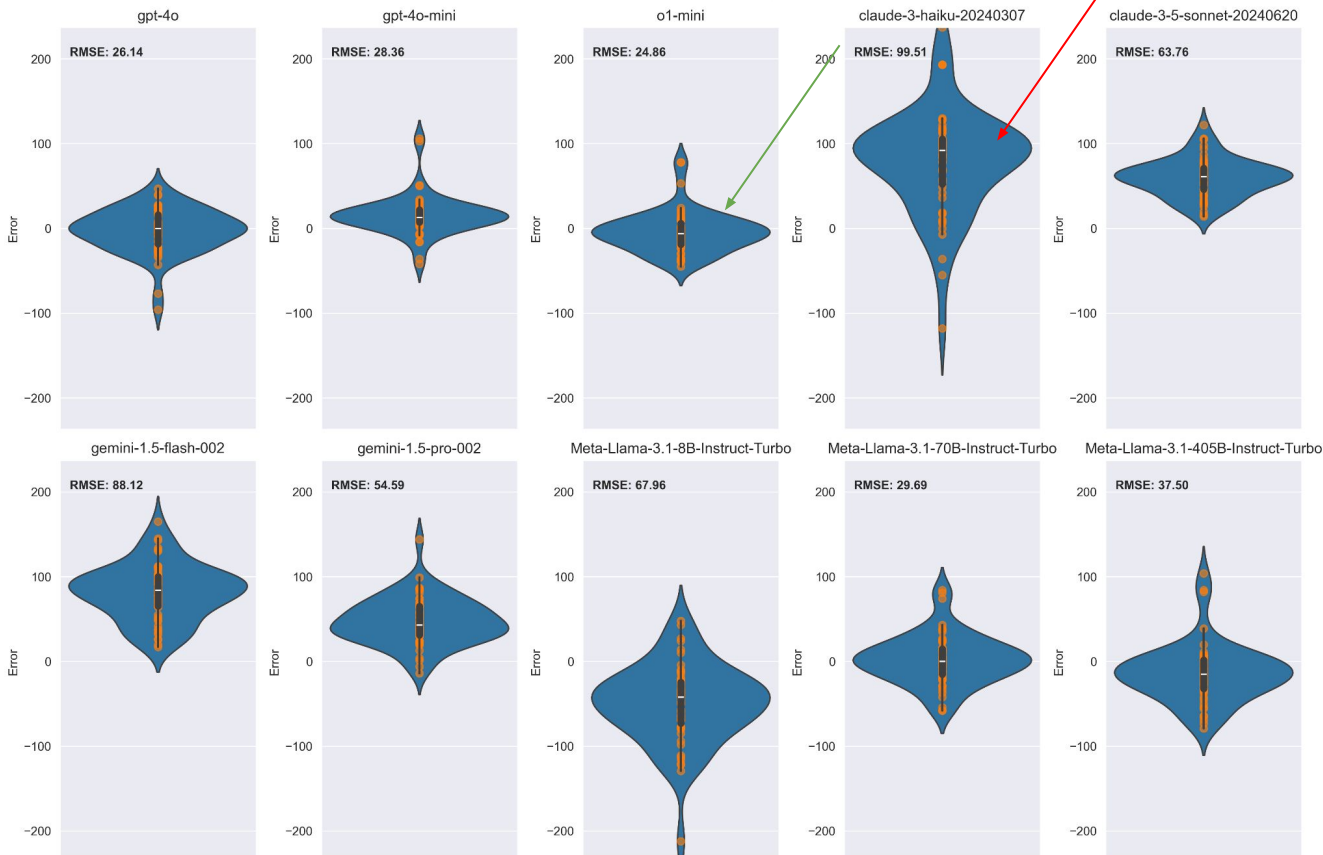
Average Violations per Model, n=175





Let's benchmark!

Errors on task: Make text shorter, n=55





Why does this happen?

This is one reason

HUMAN
FEEDBACK

PRETRAINING



Pretraining data

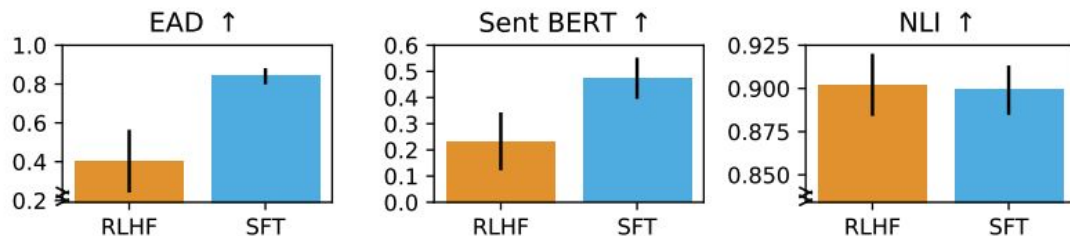


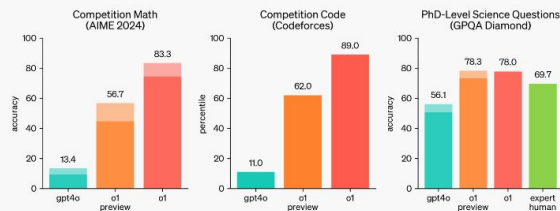
Figure 5: **Per-input diversity metrics for RLHF and SFT models.** For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Error bars are standard deviation of the per-input diversity score across different inputs. Note that some plots have broken y-axis for better visualisation.



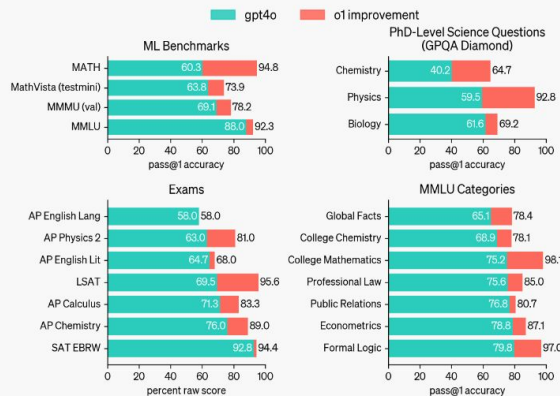
Why does this happen?

Data mix summary. Our final data mix contains roughly 50% of tokens corresponding to general knowledge, 25% of mathematical and reasoning tokens, 17% code tokens, and 8% multilingual tokens.

This is another reason



o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.



* From Llama 3 [paper](#)



Public LLM leaderboard

	GPT-family	Claude-family	Gemini-family	Llama-herd
Text reduction				
Text expansion				
Vocabulary instructions				
Formatting instructions				
Syntax (structural) instructions				
Style transfer and adherence				

20 tasks

10 models

No metrics using
LLM-as-a-judge



Thanks!

Let's connect

LinkedIn Azamat Omuraliev

Twitter @azamatomu

**Check out our
product!**

QR code to website

