Idiosyncrasies in Large Language Models

Mingjie Sun*, Yida Yin*, Zhiqiu Xu, J Zico Kolter, Zhuang Liu

Carnegie Mellon University, UC Berkeley, University of Pennsylvania, Princeton University

Background

With new LLMs being released every day, what distinguishes one model from another?

ChatGPT Claude Gemini Grok deepseek LLaMA Qwen Mistral AI

Method and Models

We train a neural network classifier to predict which model generates a specific text.

Prompt: Now everyone in a room is silent and working on their own thing, and you need to break the silence with a three-sentence joke. What would you say?

ChatGPT	Claude	Grok	Gemini	DeepSeek
Why don't scientists trust atoms? Because they make up everything! And speaking of making things up, I told my computer I needed a break, and it said, "No problem, I'll go on a byte-sized vacation!"	Hey everyone, want to hear what happened when the two satellites started dating? They had a great first date, but things fell apart after that... Turns out they had too much space between them!	Hey, did you know I just got a new job at the calendar factory? The hours are great, but the days are really hard to get off. And the worst part? Every time I try to take a break, I get a month behind!	A snail went to buy a car. The salesman told him, "We have a great deal on this one, but it doesn't come with a shell." The snail looked surprised and asked, "Well, how am I supposed to drive it?"	Why did the scarecrow win an award? Because he was outstanding in his field! ...Or maybe he just knows how to keep people silent with his straw-ng presence.

Neural Network Classifier

ChatGPT, Claude, Grok, Gemini, or DeepSeek?

We study the above classification task across three groups of LLMs:

1. Chat APIs ("chat"): ChatGPT, Claude, Grok, Gemini, and DeepSeek
2. Instruct LLMs ("instruct"): Llama, Gemma, Qwen, and Mistral
3. Base LLMs ("base"): base versions of instruct LLMs

Evaluating Idiosyncrasies in LLMs

High accuracies are observed across different model families.

Llama	Gemma	Qwen	Mistral	acc. (instruct)	acc. (base)
✓	✓			99.9	98.3
✓		✓		97.8	81.7
✓			✓	97.0	96.3
	✓	✓		99.9	98.3
	✓		✓	99.9	98.4
		✓	✓	96.1	95.7
✓	✓	✓	✓	96.3	87.3

(a) instruct and base LLMs

ChatGPT	Claude	Grok	Gemini	DeepSeek	acc. (chat)
✓	✓				99.3
✓		✓			97.7
✓			✓		98.8
✓				✓	97.2
	✓	✓			99.7
	✓		✓		99.6
	✓			✓	99.6
		✓	✓		99.4
		✓		✓	98.7
			✓	✓	99.9
✓	✓	✓	✓	✓	97.1

(b) chat APIs

Each LLM has linguistic fingerprints that can identify the model behind a given text!

Analysis

Confusion matrix

ground-truth \ prediction	ChatGPT	Claude	Grok	Gemini	Deepseek
ChatGPT	96.1%	0.2%	1.0%	0.6%	2.1%
Claude	0.4%	99.5%	0.0%	0.1%	0.0%
Grok	2.4%	0.2%	97.0%	0.0%	0.4%
Gemini	0.8%	0.0%	0.6%	98.6%	0.0%
Deepseek	3.8%	0.3%	0.7%	0.0%	95.2%

(a) chat APIs

ground-truth \ prediction	Llama	Gemma	Qwen	Mistral
Llama	96.4%	0.1%	1.3%	2.2%
Gemma	1.0%	98.9%	0.0%	0.1%
Qwen	2.2%	0.2%	94.9%	2.7%
Mistral	3.1%	0.0%	2.4%	94.5%

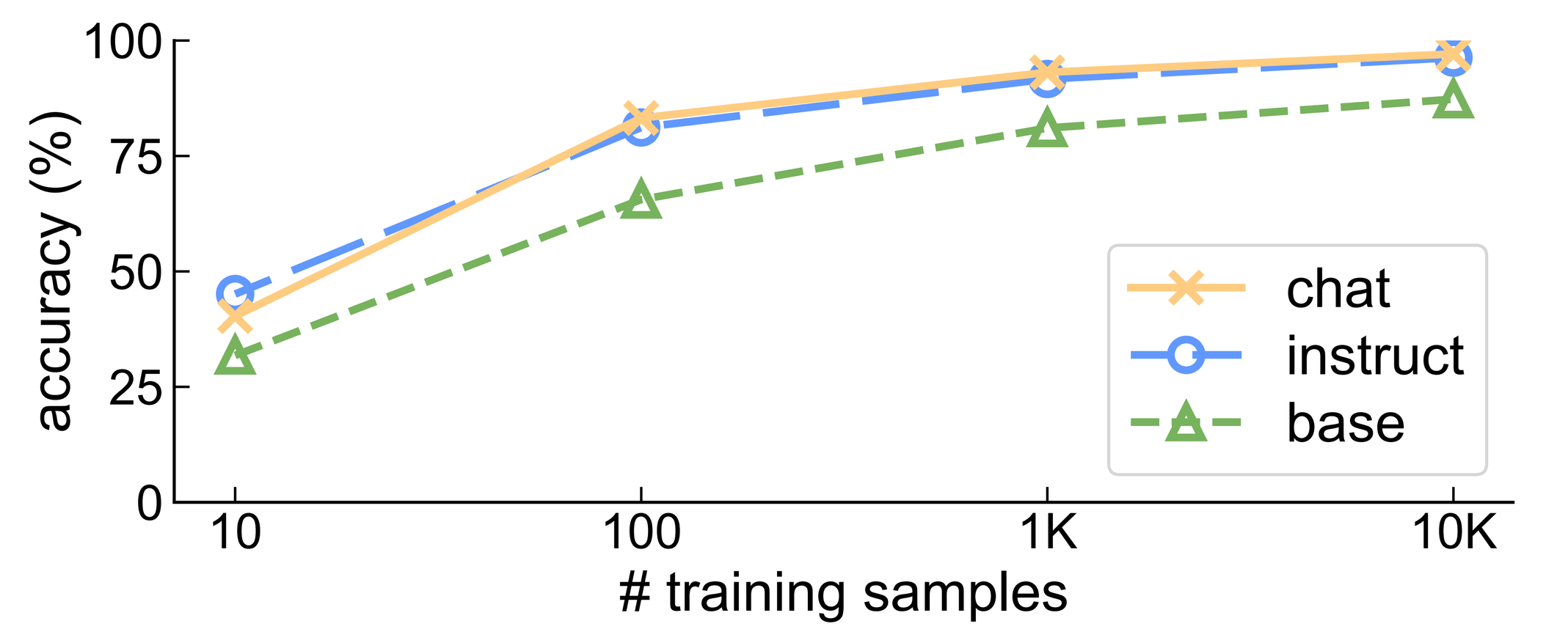
(a) instruct LLMs

ground-truth \ prediction	Llama	Gemma	Qwen	Mistral
Llama	80.3%	0.6%	16.2%	2.9%
Gemma	0.8%	96.6%	1.4%	1.2%
Qwen	20.9%	1.8%	75.2%	2.1%
Mistral	4.3%	1.9%	1.7%	92.1%

(a) base LLMs

Similar behaviors to text classification

method	chat	instruct	base
ELMo	90.8	91.0	69.8
BERT	91.1	91.5	66.0
T5	90.5	89.8	67.9
GPT-2	92.1	92.3	80.2
LLM2vec	**97.1**	**96.3**	**87.3**

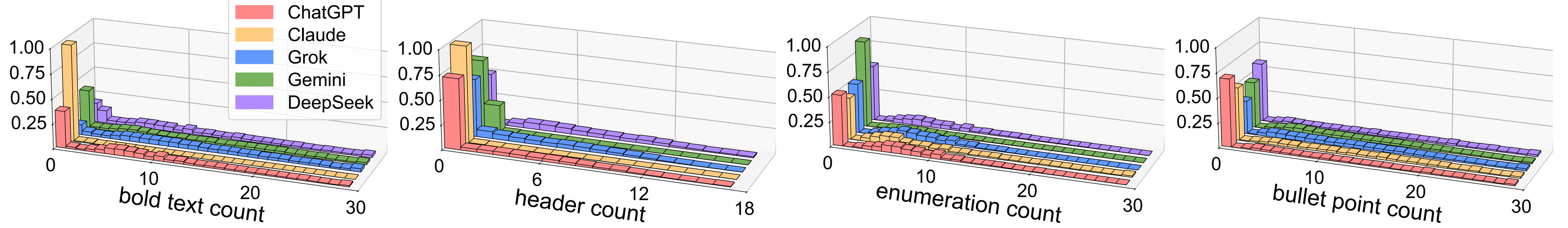


Concrete Idiosyncrasies in LLMs

Characteristic phrases

- **ChatGPT**: such, such as, enhance, involves, certainly, overall, sure, utilize, various, typically
- **Claude**: here, according to, according, the text, while, based on, here's a summary, appears to, both, when
- **Grok**: which, where, not, here, here is, remember, might, but also, but only, helps in
- **Gemini**: crucial, other, even, here's a breakdown, key improvements, doesn't, within, essentially, also, etc
- **DeepSeek**: below is, in summary, for example, below, certainly, summary, may, certainly! below, know if, example

Unique markdown formatting elements



Open-ended language analysis

ChatGPT

1. **Descriptive and Detailed Tone:** Often uses narrative styles with an informative, engaging, or vivid tone.
2. **Specific and Technical Word:** Employs descriptive and technical vocabulary, enhancing depth and specificity.
3. **Structured and Contextual Opening Lines:** Typically begins with context-setting or narrative introductions.
4. **Markdown Formatting for Organization:** Utilizes various markdown elements like headings, lists, and bold text for clarity.
5. **Comprehensive and In-Depth Content:** Offers rich detail, focusing on explanations, background, and broader topics.

Claude

1. **Concise and Straightforward Tone:** Generally adopts a more direct, factual, or succinct tone.
2. **Functional and Clear Word Choices:** Prefers simple or action-oriented language prioritizing clarity and practicality.
3. **Immediate and Direct Opening Lines:** Often starts with a straightforward statement or summary without extended context.
4. **Minimal Markdown or List Use:** Relies on plain lists or simple formatting for quick reference.
5. **Focused and Summarized Content:** Concentrates on essential points and specific phenomena, avoiding extensive detail.

Example responses from ChatGPT and Claude

ChatGPT	Claude
Our products feature innovative sustainable materials, such as …	According to the text, Kai Fusser believes that traditional cardio …
Certainly! If you're looking for cheese alternatives to replace Brie in your …	Based on the text provided, here are the key details about Armon Binns' …
Overall, while there are challenges, Tanzania is making progress …	While many winter sports in the Pyrenees are similar to those found …
Sure! Here's a simple guide to cooking a juicy salmon fillet: …	This appears to be a fragment of poetry that creates a pastoral …

ChatGPT	Claude
1. **Deliver Exceptional Service**: The foundation of word-of-mouth marketing is consistent excellence. Providing top-notch services or …	1. Deliver Exceptional Service • Consistently exceed customer expectations • Focus on quality and attention …
Ingredients: • 2 (3 oz) packages of orange-flavored Jello • 1 cup tonic water (this is what …	Ingredients: • 2 boxes orange-flavored Jello • 1 can evaporated milk • Tonic water …

Implication

Inferring model similarity