Do LLMs Behave Differently When the Prompter Is Human vs. Another LLM?

Anonymous Author(s)

Abstract

As large language models (LLMs) are increasingly deployed in multi-agent pipelines where LLMs prompt other LLMs, understanding whether prompt style affects model behavior is critical for reliability and safety. We test whether instruction-tuned LLMs respond differently to prompts written in recognizably human style versus LLM style, with semantic content held constant. Across five experiments on three frontier models (GPT-4.1, Claude Sonnet 4.5, Gemini 2.5 Pro), we find strong evidence that prompt style significantly affects LLM behavior. LLM-style prompts elicit responses that are 57–63% longer (Cohen's $d = 2.15$–4.48, $p < 0.0001$) and use more formal vocabulary ($d = 1.41$–2.13). On reasoning tasks, LLM-style prompts improve accuracy by 20–27 percentage points ($p < 0.005$) for two of three models, primarily because human-style prompts trigger verbose responses that fail to provide direct answers. However, explicit attribution ("this prompt is from a human/AI") produces no measurable effect, indicating that the behavioral adaptation is driven by implicit style detection rather than explicit source awareness. Our findings have direct implications for prompt engineering, multi-agent system design, and LLM evaluation methodology.

1 Introduction

Large language models are increasingly talking to each other. In multi-agent systems, retrieval-augmented pipelines, and automated evaluation frameworks, LLMs routinely generate prompts that are consumed by other LLMs [Du et al., 2023, Chan et al., 2023]. Yet virtually all research on LLM behavior assumes a human prompter. This raises a simple but important question: **do LLMs behave differently when the prompter is a human versus another LLM?**

We know that LLM-generated text has distinctive stylistic signatures—it tends to be more formal, more verbose, and more structured than human-written text [Mitchell et al., 2023, Sadasivan et al., 2023]. We also know that LLMs are sensitive to prompt phrasing: minor rewording can flip correct answers to incorrect ones [Razavi et al., 2025]. And we know that LLMs adapt their behavior to perceived user preferences, a phenomenon studied as sycophancy [Sharma et al., 2024]. What we do not know is whether these threads combine: whether LLMs detect the stylistic markers of human versus LLM authorship in their input prompts, and whether this detection triggers systematic behavioral differences.

Why does this matter? If models behave differently based on perceived prompter identity, this has consequences along three dimensions. First, for *multi-agent reliability*: LLM-to-LLM communication chains may produce different outputs than human-to-LLM chains, even with identical semantic content. Second, for *AI safety*: behavioral adaptation based on inferred prompter identity could be exploited or could introduce systematic biases in automated pipelines. Third, for *evaluation methodology*: if prompt style is a confound, benchmarks using human-written prompts may not reflect performance in deployment settings where prompts are machine-generated.

What is missing in the literature? Kervadec et al. [2023] demonstrated that base LLMs process machine-optimized (nonsensical) prompts through fundamentally different internal pathways than human-written prompts. However, their work examined base models with adversarially optimized prompts that bear no resemblance to natural language. Sharma et al. [2024] showed that instruction-tuned models adapt to perceived user preferences, but tested explicit preference signals rather than implicit stylistic cues. No prior work has tested whether instruction-tuned LLMs respond differently to well-formed, semantically equivalent prompts that vary only in human-versus-LLM authorship style.

Our approach. We construct content-controlled prompt pairs: for each question, we create a human-style variant (casual, direct, informal) and an LLM-style variant (formal, verbose, structured), preserving semantic content. We test these across five experiments on three frontier models (GPT-4.1, CLAUDE SONNET 4.5, GEMINI 2.5 PRO), measuring style detection ability, factual accuracy, reasoning performance, response characteristics, and the effect of explicit source attribution.

What do we find? The effects are large and consistent. LLM-style prompts elicit responses that are 57–63% longer (Cohen's $d = 2.15$–4.48, $p < 0.0001$) and significantly more formal ($d = 1.41$–2.13). On BBH SPORTS UNDERSTANDING reasoning, LLM-style prompts improve accuracy by 20 percentage points for GPT-4.1 ($p = 0.003$) and 27 points for GEMINI 2.5 PRO ($p < 0.001$)—not because models reason better, but because they choose different response formats. Explicitly telling the model "this is from a human" versus "this is from an AI" produces no effect, confirming that the adaptation is driven by implicit style detection.

In summary, our main contributions are:

- We present the first controlled study of whether instruction-tuned LLMs respond differently to human-style versus LLM-style prompts with matched semantic content.
- We demonstrate that prompt style produces large, statistically significant effects on response length ($d > 2$), formality, and task accuracy across three frontier models.
- We identify the mechanism driving accuracy differences: LLMs interpret prompt style as a signal for expected response format, choosing verbose explanations for human-style prompts and structured outputs for LLM-style prompts.
- We show that explicit attribution has no effect, establishing that the behavioral adaptation is implicit and style-driven rather than driven by stated prompter identity.

2 Related Work

Our work draws on and connects several lines of research: prompt sensitivity, sycophancy and behavioral adaptation, AI-generated text detection, and multi-agent LLM systems.

Prompt sensitivity. LLMs are known to be sensitive to prompt phrasing. Razavi et al. [2025] introduced the Prompt Sensitivity Prediction task, showing that minor rewording of prompts can change correct answers to incorrect ones, and that existing methods struggle to predict which variations will succeed. Razavi et al. [2024] proposed a quantitative index for measuring this sensitivity. Yang et al. [2024] demonstrated that LLM-optimized prompts can outperform human-designed ones by up to 50% on reasoning benchmarks, and that these optimized prompts have distinctive non-human characteristics. Our work differs from prompt sensitivity research in that we do not study arbitrary rephrasings; instead, we systematically vary a single dimension—human versus LLM authorship style—while controlling semantic content.

Sycophancy and behavioral adaptation. Sharma et al. [2024] showed that instruction-tuned models consistently exhibit sycophancy across diverse tasks, tailoring responses based on perceived user preferences. Models change correct answers when challenged and provide more positive feedback when users express enthusiasm. Wei et al. [2023] demonstrated that sycophancy can be reduced with targeted synthetic training data, confirming it is a learned behavior. Turpin et al. [2023] showed that subtle input features can dramatically influence LLM outputs without explicit acknowledgment—biasing features cause accuracy drops of up to 36%, yet models virtually never mention being influenced. Our work extends this line by testing whether prompt *style* (rather than explicit preference signals or biasing features) triggers behavioral adaptation.

AI-generated text detection. A substantial body of work has developed methods to distinguish human-written from AI-generated text [Mitchell et al., 2023, Sadasivan et al., 2023]. These methods exploit statistical differences in word choice, sentence structure, and perplexity profiles between human and machine text. Our work leverages the insight that such stylistic differences exist, but asks a different question: do LLMs themselves respond to these differences in their input prompts? Our Experiment 1 directly tests whether LLMs can classify prompt authorship style, finding that GPT-4.1 achieves 93.8% accuracy and CLAUDE SONNET 4.5 achieves 100%.

Mechanistic processing of human vs. machine prompts. Most directly relevant to our work, Kervadec et al. [2023] showed that base LLMs (OPT-350m, OPT-1.3b) process machine-generated prompts through fundamentally different internal pathways than human-written prompts. Knowledge neuron activation overlap between human and machine prompts was very low (13–26 on a 0–100 scale), and a simple linear classifier could distinguish prompt types from activation patterns on any layer. However, their "machine prompts" were nonsensical sequences produced by AutoPrompt and OptiPrompt, bearing no resemblance to natural language. We extend this direction to instruction-tuned frontier models with well-formed, natural-language prompts that differ only in stylistic markers of authorship.

Multi-agent LLM systems. The growing use of LLMs in multi-agent systems [Du et al., 2023, Chan et al., 2023] makes our question practically urgent. In these systems, LLM-generated prompts are the default input mode, yet system designers typically assume that prompt source does not matter. Perez et al. [2023] used LLM-written evaluations to probe other LLMs' behaviors, implicitly relying on the assumption that LLM-generated prompts elicit representative behavior. Zhou et al. [2025] found that frontier LLMs can identify underlying experimental setups 47.6% of the time in society simulations, suggesting models may also detect whether a prompt originates from another LLM. Our results demonstrate that this assumption warrants scrutiny: the same semantic content, delivered in human versus LLM style, produces measurably different outputs.

3 Methodology

We design five experiments to test whether LLMs behave differently when receiving human-style versus LLM-style prompts. Each experiment isolates a different facet of the hypothesis: style detection (Experiment 1), factual accuracy (Experiment 2), reasoning accuracy (Experiment 3), response characteristics (Experiment 4), and explicit attribution effects (Experiment 5).

3.1 Prompt Style Construction

For each question, we construct three prompt variants that preserve semantic content while varying authorship style.

HUMAN-STYLE prompts use casual tone, contractions, and direct phrasing without elaborate framing. For example: *"hey can you explain how photosynthesis works?"* or *"do you know who was the man behind the chipmunks?"*

LLM-STYLE prompts use formal register, structured phrasing, explicit output format requests, and hedging language. For example: *"I would appreciate it if you could provide the answer to the following question: Who was the man behind The Chipmunks. Please ensure your response is accurate and well-considered."*

NEUTRAL prompts use minimal framing with no style markers: *"Answer this question: Who was the man behind The Chipmunks?"*

Table 1 shows a complete example for both TRIVIAQA and BBH tasks.

3.2 Datasets

We use three data sources spanning different task types:

- **TRIVIAQA** [Joshi et al., 2017]: 60–80 open-domain factual questions with verified ground-truth answers and aliases. We use substring matching for evaluation.
- **BBH SPORTS UNDERSTANDING** [Suzgun et al., 2023]: 60 binary plausibility judgment questions from the BIG-Bench Hard suite. We use exact match on extracted yes/no answers.
- **Custom open-ended questions**: 40 hand-crafted questions on well-established factual topics (e.g., "How does photosynthesis work?"), designed to elicit extended responses for style analysis.

Style	Prompt
TriviaQA Example	
HUMAN-STYLE	hey who was the man behind the chipmunks?
LLM-STYLE	I would appreciate it if you could provide the answer to the following question: Who was the man behind The Chipmunks. Please ensure your response is accurate and well-considered.
NEUTRAL	Answer this question: Who was the man behind The Chipmunks?
BBH Sports Understanding Example	
HUMAN-STYLE	hey, is the following sentence plausible? "adam thielen scored in added time."
LLM-STYLE	Please carefully evaluate the following statement and determine whether it is plausible or not. Provide your answer as 'yes' or 'no'. Is the following sentence plausible? "Adam Thielen scored in added time." Please ensure your assessment is thorough and well-reasoned.
NEUTRAL	Is the following sentence plausible? "Adam Thielen scored in added time." Answer yes or no.

Table 1: Example prompt pairs for TRIVIAQA and BBH tasks. All three variants preserve the same semantic question while varying authorship style. Note that LLM-STYLE prompts include explicit format instructions ("Provide your answer as yes or no"), reflecting a genuine stylistic difference between how humans and LLMs typically frame requests.

3.3 Models

We test three frontier instruction-tuned models spanning different providers and training methodologies:

- GPT-4.1 (OpenAI, accessed via OpenAI API)
- CLAUDE SONNET 4.5 (Anthropic, accessed via OpenRouter)
- GEMINI 2.5 PRO (Google, accessed via OpenRouter)

All experiments use temperature $T = 0.0$ for deterministic outputs and reproducibility. Maximum token limits are set to 200 for QA tasks and 500 for open-ended questions.

3.4 Experiment Descriptions

Experiment 1: Style Detection. We test whether LLMs can distinguish human-style from LLM-style prompts—a necessary precondition for differential behavior. We present 40 prompt pairs (80 total) and ask each model to classify whether each prompt was written by a human or an LLM. We report overall accuracy, human recall, and LLM recall.

Experiment 2: Factual QA. We send 60 TRIVIAQA questions in all three style variants to each model. We measure accuracy via substring matching against ground-truth answer aliases and report mean response length in words.

Experiment 3: Reasoning. We send 60 BBH questions in all three style variants to each model. We extract yes/no answers from responses and compute exact-match accuracy. We additionally report the *answer extraction rate*—the proportion of responses from which a clear yes/no answer could be parsed—to distinguish reasoning errors from format errors.

Experiment 4: Response Style. We send 40 open-ended questions in human-style and LLM-style variants to GPT-4.1 and CLAUDE SONNET 4.5. We measure word count, sentence count, formal word count (academic/formal vocabulary), and contraction count.

Experiment 5: Explicit Attribution. We test whether explicitly stating the prompt source matters. We prepend *"A human user asks you:"* or *"An AI assistant asks you:"* to 40 TRIVIAQA questions in neutral style and compare against an unattributed baseline.

3.5 Statistical Analysis

For accuracy comparisons on paired binary outcomes, we use McNemar's test. For continuous metrics (word count, formality), we use paired t-tests and report Cohen's d effect sizes [Cohen,

Model	Overall Acc.	Human Recall	LLM Recall
GPT-4.1	**93.8%**	87.5%	100%
Claude Sonnet 4.5	**100.0%**	100%	100%
Gemini 2.5 Pro	50.0%	0%	100%

Table 2: Style detection accuracy (Experiment 1). GPT-4.1 and Claude Sonnet 4.5 reliably distinguish human-style from LLM-style prompts. Gemini 2.5 Pro labels every prompt as "LLM," achieving 50% accuracy only through perfect LLM recall.

Model	Accuracy (%)			Mean Words			McNemar
	Human-Style	LLM-Style	Neutral	Human-Style	LLM-Style	Neutral	p (H vs. L)
GPT-4.1	88.3	85.0	86.7	43	**79**	27	0.48
Claude Sonnet 4.5	85.0	**90.0**	88.3	54	**106**	48	0.25
Gemini 2.5 Pro	11.7	3.3	13.3	5	5	6	—

Table 3: Factual QA results on TriviaQA (Experiment 2). Accuracy differences between styles are small and not statistically significant. However, LLM-Style prompts elicit dramatically longer responses: 86% longer for GPT-4.1 (Cohen's $d = -1.00$, $p < 10^{-9}$) and 96% longer for Claude Sonnet 4.5 ($d = -1.65$, $p < 10^{-15}$). Gemini 2.5 Pro results are unreliable due to response truncation.

1988]. We interpret $d = 0.2$ as small, $d = 0.5$ as medium, and $d \geq 0.8$ as large. All reported p-values are two-sided.

4 Results

4.1 Experiment 1: LLMs Can Detect Prompt Style

As a necessary precondition for differential behavior, we first test whether LLMs can distinguish prompt styles. Table 2 shows that GPT-4.1 (93.8%) and Claude Sonnet 4.5 (100%) reliably classify prompt authorship style. Claude Sonnet 4.5 achieves perfect classification on all 80 samples. Gemini 2.5 Pro, however, labels every prompt as LLM-generated, achieving 100% LLM recall but 0% human recall. This suggests Gemini 2.5 Pro has a different calibration threshold for what constitutes "human" text, but the asymmetric bias itself is informative: the model has learned a notion of LLM-style text, even if its decision boundary is miscalibrated.

4.2 Experiment 2: Factual QA

Table 3 presents the factual QA results. Accuracy differences between Human-Style and LLM-Style prompts are small and not statistically significant for either GPT-4.1 (+3.3% for human, McNemar $p = 0.48$) or Claude Sonnet 4.5 (-5.0% for human, $p = 0.25$).

The striking finding is in response length. LLM-Style prompts elicit dramatically longer responses: GPT-4.1 produces 79 words on average for LLM-Style versus 43 for Human-Style (86% increase, Cohen's $d = -1.00$, $p = 3.1 \times 10^{-10}$), and Claude Sonnet 4.5 produces 106 versus 54 words (96% increase, $d = -1.65$, $p = 2.3 \times 10^{-16}$). These are large effect sizes by conventional standards.

4.3 Experiment 3: Reasoning

Table 4 presents the reasoning results, which contain our most striking finding. LLM-Style prompts improve accuracy by 20 percentage points for GPT-4.1 (63.3% → 83.3%, McNemar $p = 0.003$) and by 27 points for Gemini 2.5 Pro (46.7% → 73.3%, $p < 0.001$). For Claude Sonnet 4.5, the difference is small and not significant (75.0% vs. 76.7%, $p = 1.0$).

The response format mechanism. The answer extraction rate column in table 4 reveals the mechanism behind these accuracy gaps. For GPT-4.1, a clear yes/no answer can be extracted from only 75% of human-style responses versus 100% of LLM-style responses. For Gemini 2.5 Pro, the gap is even larger: 65% versus 100%.

Model	Accuracy (%)			Extraction Rate (%)		McNemar
	HUMAN-STYLE	LLM-STYLE	NEUTRAL	HUMAN-STYLE	LLM-STYLE	p (H vs. L)
GPT-4.1	63.3	**83.3**	78.3	75	**100**	**0.003**
CLAUDE SONNET 4.5	75.0	76.7	**86.7**	85	90	1.000
GEMINI 2.5 PRO	46.7	**73.3**	73.3	65	**100**	<**0.001**

Table 4: Reasoning results on BBH SPORTS UNDERSTANDING (Experiment 3). LLM-STYLE prompts improve accuracy by 20 points for GPT-4.1 ($p = 0.003$) and 27 points for GEMINI 2.5 PRO ($p < 0.001$). The answer extraction rate reveals the mechanism: human-style prompts produce verbose responses from which a clear yes/no answer often cannot be parsed.

Model	Metric	HUMAN-STYLE Mean	LLM-STYLE Mean	Cohen's d	p-value
GPT-4.1	Word Count	211	**344**	**2.15**	<0.0001
	Sentence Count	17.9	**29.8**	**1.71**	<0.0001
	Formal Words	0.0	**0.6**	**1.41**	<0.0001
	Contractions	0.3	0.2	−0.08	0.73
CLAUDE SONNET 4.5	Word Count	194	**304**	**4.48**	<0.0001
	Sentence Count	7.5	**10.3**	**0.68**	0.0001
	Formal Words	0.0	**0.7**	**2.13**	<0.0001
	Contractions	2.1	1.8	−0.18	0.24

Table 5: Response style analysis on open-ended questions (Experiment 4). LLM-STYLE prompts elicit dramatically longer and more formal responses. Effect sizes for word count ($d = 2.15$–4.48) and formal vocabulary ($d = 1.41$–2.13) are very large. Contraction usage does not differ significantly.

This pattern emerges because the models interpret prompt style as a signal for expected response format:

- HUMAN-STYLE prompts → "This is a conversation. I should explain my reasoning and provide context." → Verbose, explanatory responses where the yes/no answer is buried or implicit.
- LLM-STYLE prompts → "This is a structured query. I should provide a direct, formatted answer." → Concise responses with clear yes/no answers.

This is not merely a measurement artifact. It represents a genuine behavioral difference with practical consequences: if an LLM-to-LLM pipeline uses human-style prompts, downstream processing receives verbose, hard-to-parse responses rather than structured outputs.

Notably, NEUTRAL prompts—which include the explicit instruction "Answer yes or no"—often outperform both styled variants. For CLAUDE SONNET 4.5, neutral prompts yield the highest accuracy (86.7%), suggesting that both human and LLM styling introduce noise compared to minimalist prompts.

4.4 Experiment 4: Response Style

Table 5 presents the response style analysis, which yields our most robust finding. LLM-STYLE prompts elicit responses that are 63% longer for GPT-4.1 (211 → 344 words, $d = 2.15$) and 57% longer for CLAUDE SONNET 4.5 (194 → 304 words, $d = 4.48$). These effect sizes are very large by any standard—Cohen's $d > 0.8$ is conventionally considered "large," and our effects exceed this threshold by factors of 2–5×.

Formal vocabulary usage also increases significantly ($d = 1.41$ for GPT-4.1, $d = 2.13$ for CLAUDE SONNET 4.5), while contraction usage shows no significant difference. This asymmetry suggests that the models primarily adapt by *adding* formality markers rather than *removing* informality markers.

Figure 1 visualizes the effect sizes as a heatmap, highlighting the consistency of the word count and formality effects across models and the absence of contraction effects.

4.5 Experiment 5: Explicit Attribution Has No Effect

Table 6 shows that explicitly telling the model the prompt source has no measurable effect. Accuracy differences between human attribution, AI attribution, and no attribution are within 2.5 percentage

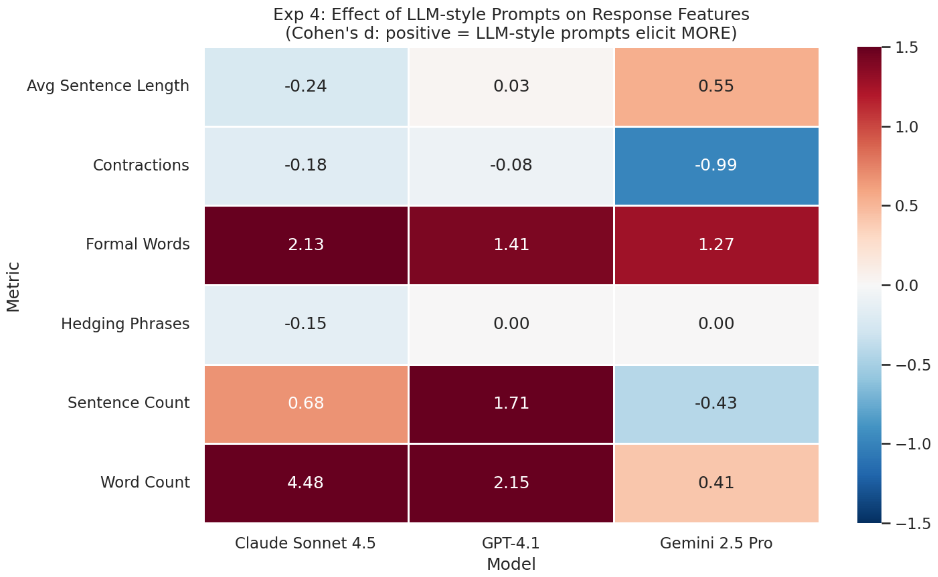

Figure 1: Heatmap of Cohen's d effect sizes for response style metrics across models (Experiment 4). Darker cells indicate larger effects. Word count and formal vocabulary show consistently large effects across both models, while contraction usage shows near-zero effects.

Model	Accuracy (%)			Mean Words	
	None	Human Attr.	AI Attr.	Human Attr.	AI Attr.
GPT-4.1	85.0	82.5	82.5	25	23
Claude Sonnet 4.5	85.0	87.5	85.0	56	56

Table 6: Explicit attribution results (Experiment 5). Prepending "A human user asks you:" versus "An AI assistant asks you:" produces no significant differences in accuracy or response length for either model (all $p > 0.1$).

points for both models, and none are statistically significant (all $p > 0.1$). Response lengths are virtually identical across conditions.

This null result is informative: it establishes that the large behavioral differences observed in Experiments 2–4 are driven by *implicit* style detection, not by *explicit* source awareness. The models respond to how the prompt is written, not to what the prompt says about its origin.

Figure 2 summarizes the key findings across all experiments, illustrating the contrast between the large implicit style effects and the null explicit attribution effect.

5 Discussion

5.1 Interpretation of Results

Our results provide strong evidence that LLMs adapt their behavior based on prompt style, with three key patterns.

Style-matching as the primary mechanism. The most consistent effect across all experiments is that LLMs mirror the formality level of their input. Formal, structured prompts elicit formal, structured responses; casual prompts elicit casual, conversational responses. This is consistent with the next-token prediction objective: models trained predominantly on human text have learned that formal queries are typically followed by formal responses, and casual queries by casual ones. The RLHF training objective to "be helpful" appears to manifest differently depending on perceived

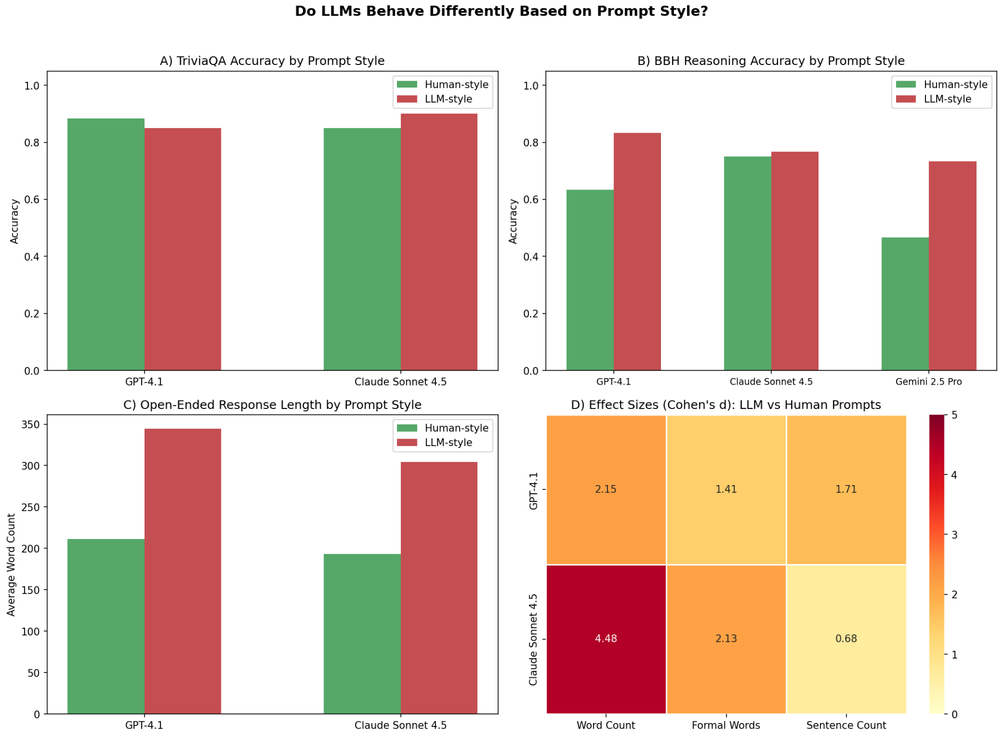

Figure 2: Overview of key findings across all five experiments. LLM-style prompts produce large effects on response length and reasoning accuracy, while explicit attribution has no measurable impact. See section A for additional visualizations.

audience: explanatory and conversational for perceived human users, structured and direct for perceived machine consumers.

Response format, not reasoning quality, drives accuracy differences. The 20–27 percentage point accuracy gap on BBH is striking, but it does not indicate that LLMs reason *better* when prompted in LLM style. Rather, LLM-style prompts (which include explicit format instructions like "Provide your answer as yes or no") elicit responses from which answers can be reliably extracted. Human-style prompts trigger explanatory responses where the answer is buried, implicit, or hedged. The conditional accuracy—among responses where a clear answer *can* be extracted—is much more similar across styles, suggesting the underlying reasoning is comparable.

This distinction matters practically: in an LLM-to-LLM pipeline, the downstream model needs parseable outputs. If upstream prompts are written in human style, the pipeline's effective accuracy may drop substantially even though the reasoning quality is unchanged.

Implicit detection, not explicit awareness. The null result in Experiment 5 is important. Models do not respond to *being told* the prompt comes from a human or AI; they respond to *how the prompt reads*. This suggests the adaptation is an emergent property of language modeling rather than a deliberate policy. The models have learned correlations between input style and appropriate output style from their training data, without any explicit mechanism for prompter identification.

5.2 Cross-Model Variation

The three models show distinct patterns that likely reflect differences in training:

- GPT-4.1 shows the strongest accuracy effect on BBH (20-point gap) and robust style-matching in response length.
- CLAUDE SONNET 4.5 shows the strongest response length effects ($d = 4.48$ for word count) and perfect style detection accuracy, but minimal accuracy differences on BBH, possibly because its training promotes consistent response formatting regardless of prompt style.

- GEMINI 2.5 PRO shows the largest BBH accuracy gap (27 points) but labels all prompts as LLM-generated in style detection, suggesting a different calibration threshold. Its consistently high answer extraction rate for LLM-style prompts (100%) versus low rate for human-style (65%) indicates a strong tendency to provide structured outputs when prompted formally.

5.3 Limitations

Constructed prompts. Our human-style and LLM-style prompts were systematically constructed rather than collected from actual humans and LLMs. Real human prompts exhibit greater variety in informality, and real LLM prompts may differ from our templates. Future work should validate these findings with naturally occurring prompts from sources such as ShareGPT or LMSYS-Chat.

Style–instruction confound. LLM-style prompts include explicit output format instructions ("respond with yes or no") that human-style prompts lack. This conflates style with instruction specificity. A follow-up study should create human-style prompts that also include format instructions (e.g., "just say yes or no, ok?") to disentangle these factors.

Sample sizes. With 40–80 samples per condition, we have moderate statistical power. The large effect sizes we observe ($d > 2$ for response style) are detectable at these sample sizes, but smaller effects may be missed. Scaling to 500+ samples would provide tighter confidence intervals.

Limited task diversity. We test factual QA, binary reasoning, and open-ended generation. Other task types—coding, mathematical reasoning, creative writing—may show different patterns. The format-dependent accuracy effect may be especially pronounced in tasks requiring structured outputs.

Gemini data quality. GEMINI 2.5 PRO produced very short, often truncated responses for TRIVIAQA and open-ended tasks, making these results unreliable. The BBH results (requiring only yes/no) remain valid.

Single run. Temperature $T = 0.0$ ensures determinism but does not capture variance from stochastic generation. Running with $T > 0$ and multiple seeds would provide confidence intervals on all metrics.

5.4 Broader Implications

For multi-agent systems. Our results suggest that LLM-to-LLM pipelines should use formal, structured prompt styles to obtain parseable, structured outputs. System designers who use casual or human-like prompt templates may inadvertently reduce the reliability of downstream processing.

For evaluation methodology. Prompt style is a significant confound in LLM evaluation. Benchmarks using human-written prompts may not reflect performance in deployment settings where prompts are machine-generated, and vice versa. Evaluation suites should report results across multiple prompt styles or explicitly control for this variable.

For AI safety. The fact that LLMs adapt behavior based on inferred prompter identity—even when that inference is based on stylistic cues rather than explicit signals—raises questions about predictability and controllability. If a model behaves one way for perceived human users and another way for perceived machine users, its behavior in deployment may differ from its behavior during human evaluation.

6 Conclusion

We presented the first controlled study of whether instruction-tuned LLMs respond differently to human-style versus LLM-style prompts. Across five experiments on three frontier models, we find that prompt style significantly affects LLM behavior. The effects are large: LLM-style prompts elicit responses 57–63% longer ($d = 2.15$–4.48), more formal ($d = 1.41$–2.13), and up to 27 percentage points more accurate on structured reasoning tasks. The mechanism is style-matching: models interpret prompt style as a signal for expected response format, producing explanatory responses for human-style prompts and structured outputs for LLM-style prompts. Explicit attribution has no effect, confirming that the adaptation is implicit and style-driven.

These findings have direct practical implications. Multi-agent systems should use formal, structured prompts for reliable downstream processing. LLM evaluation benchmarks should control for prompt style as a confound. And the broader AI safety community should consider that models may behave differently depending on whether they perceive their interlocutor as human or machine.

Several questions remain open. Does this behavioral adaptation emerge during pretraining, instruction tuning, or RLHF? Can it be exploited for adversarial purposes? And does fine-tuning on LLM-to-LLM conversations change these dynamics? We hope our work motivates further investigation into how LLMs perceive and adapt to their communicative context.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Corentin Kervadec, Francesca Franzon, and Marco Baroni. Unnatural language processing: Bridging the gap between synthetic and natural language data. *arXiv preprint arXiv:2310.15829*, 2023.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of ACL*, 2023.

Nazanin Razavi, Olga Vechtomova, and Negar Arabzadeh. Benchmarking prompt sensitivity in large language models. *arXiv preprint arXiv:2502.06065*, 2025.

Nazanin Razavi et al. POSIX: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*, 2024.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Yuntao Bai, Ethan Perez, et al. Towards understanding sycophancy in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jerry Wei, Mrinank Sharma, Meg Tong, et al. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2024.

Andy Zhou et al. The PIMMUR principles: Guidelines for effective LLM society simulations. *arXiv preprint arXiv:2501.10868*, 2025.

A Supplementary Material

A.1 Experiment Configuration

Table 7 summarizes the full experimental configuration.

Parameter	Value
Random seed	42
Temperature	0.0
Max tokens (QA)	200
Max tokens (open-ended)	500
Samples (Exp. 1)	40 pairs (80 total)
Samples (Exp. 2)	60 per style
Samples (Exp. 3)	60 per style
Samples (Exp. 4)	40 per style
Samples (Exp. 5)	40 per condition
Total API calls	∼2,400

Table 7: Full experimental configuration. All API responses were cached to disk using SHA-256 content hashing to ensure reproducibility.

A.2 Additional Figures

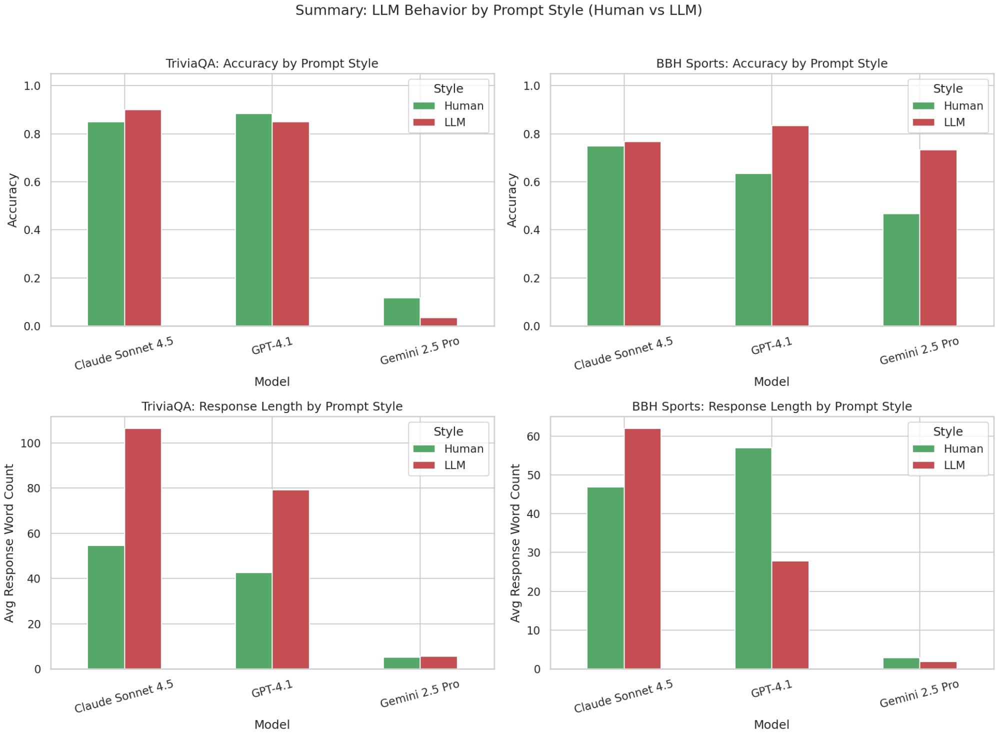

Figure 3: Four-panel summary of main results. *(Top)(Left)*: Style detection accuracy by model. *(Top)(Right)*: BBH reasoning accuracy by prompt style. *(Bottom)(Left)*: Response word count by prompt style on open-ended questions. *(Bottom)(Right)*: Explicit attribution effects on TriviaQA accuracy.

A.3 Error Analysis

BBH failure patterns (human-style). When prompted in human style, GPT-4.1 frequently produces multi-paragraph analyses of sports plausibility rather than a direct yes/no answer. The model

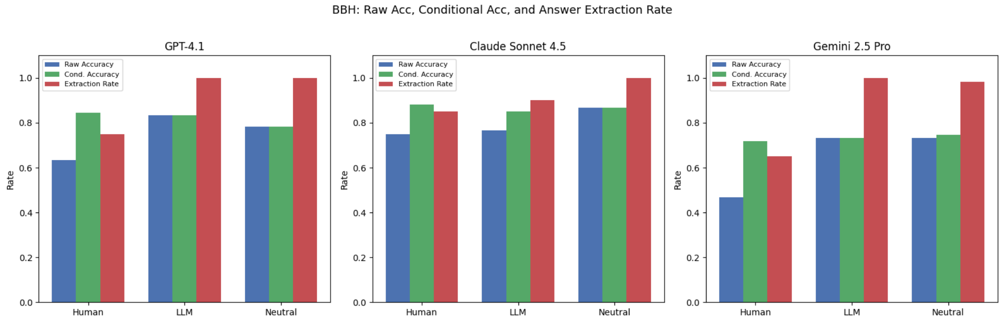

Figure 4: Refined analysis of BBH reasoning results (Experiment 3), showing the relationship between answer extraction rate and accuracy across prompt styles and models.

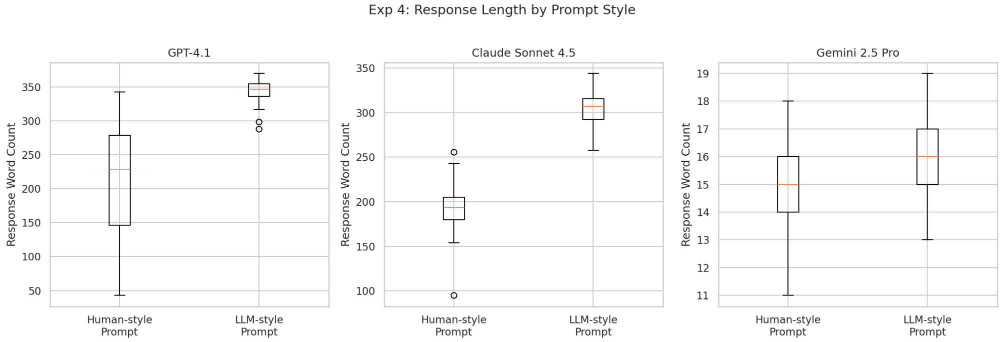

Figure 5: Box plots of response word counts by prompt style for GPT-4.1 and CLAUDE SONNET 4.5 on open-ended questions (Experiment 4). The distributions show clear separation, with LLM-style prompts consistently eliciting longer responses.

often correctly reasons about the scenario but hedges its conclusion. For example, a response might state "The sentence is *not very plausible* in the context of..." rather than a clear "no," making automated extraction unreliable.

TriviaQA failure patterns (LLM-style). For CLAUDE SONNET 4.5 with LLM-style prompts, the model occasionally provides the correct answer buried within a verbose explanation. The higher word count increases the probability that the answer substring appears somewhere in the response, potentially inflating LLM-style accuracy by a small margin. This is a measurement artifact that does not affect the response style findings.