Style Mirroring in Large Language Models: How Prompt Style Affects Response Behavior

Anonymous Author(s)

Abstract

Large Language Models (LLMs) are increasingly deployed in systems where they receive prompts from both humans and other LLMs. We investigate whether LLMs exhibit different response behaviors when presented with prompts written in a human style versus prompts written in a characteristic LLM style, while controlling for semantic content. We construct a dataset of 50 diverse questions, each paired with stylistically manipulated prompts—one informal and colloquial (human-style) and one formal and structured (LLM-style)—and query GPT-4.1-mini and Claude Sonnet 4 with both versions. Our analysis reveals that LLMs significantly alter their response behavior based on prompt style: responses to LLM-style prompts are 66% longer on average (Cohen's $d = 2.07$, $p < 0.0001$), use 120% more bullet points ($d = 1.44$), exhibit higher reading difficulty (Flesch-Kincaid Grade increases from 13.4 to 19.6), and show reduced vocabulary diversity. These effects are consistent across both model families tested. We term this phenomenon *style mirroring*—the tendency of LLMs to adapt their output style to match the stylistic characteristics of the input prompt, analogous to conversational accommodation in human communication. Our findings have implications for prompt engineering, multi-agent system design, and AI safety evaluation, suggesting that stylistic framing is an important and underexplored dimension of LLM behavior.

1 Introduction

Large Language Models (LLMs) have become fundamental infrastructure for a wide range of applications, from conversational assistants to complex multi-agent systems [Brown et al., 2020, OpenAI, 2023, Anthropic, 2024]. As these systems proliferate, an important question emerges: do LLMs behave differently depending on *who* they perceive as the source of a prompt? In particular, with the rise of LLM-to-LLM communication in agentic systems [Guo et al., 2024], understanding whether LLMs respond differently to prompts that appear to originate from humans versus other LLMs becomes critical for system design and alignment.

Prior work has established that human-written and LLM-generated text exhibit distinctive stylistic differences. Human text tends to be shorter, more colloquial, emotionally expressive, and exhibits higher vocabulary diversity, while LLM text is typically longer, more formal, logically structured, and uses a smaller vocabulary with lower word density [Guo et al., 2023, Dugan et al., 2024]. These differences are robust enough that classifiers can reliably distinguish between human and machine-generated text. A natural question arises: if stylistic features can signal the origin of text, do LLMs implicitly respond to these signals by adapting their behavior?

Research on LLM sycophancy [Sharma et al., 2024] demonstrates that models trained with human feedback tend to adapt their responses to match perceived user preferences. Similarly, work on prompt susceptibility [Anagnostidis and Bulian, 2024] shows that LLMs alter their behavior based on perceived authority and confidence signals in prompts. However, no prior work has directly in-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

vestigated whether the stylistic characteristics associated with human versus LLM authorship affect model behavior when semantic content is controlled.

Research Question. We investigate: *Do LLMs exhibit different behaviors when presented with prompts written in human style versus prompts written in LLM style, when the semantic content is held constant?*

Our Contribution. We present the first controlled study of how prompt style affects LLM response behavior. Our contributions are:

1. **Novel experimental paradigm.** We design a methodology to create semantically equivalent prompt pairs that differ only in stylistic features characteristic of human versus LLM writing, based on empirically-grounded characteristics from the detection literature.
2. **Strong empirical evidence for style mirroring.** We demonstrate that LLMs significantly alter their response behavior based on prompt style, with large effect sizes: responses to formal LLM-style prompts are 66% longer (Cohen's $d = 2.07$), use 120% more bullet points ($d = 1.44$), and exhibit substantially higher reading difficulty.
3. **Cross-model validation.** We show that these effects are consistent across two major model families (GPT-4 and Claude), suggesting a general phenomenon rather than model-specific behavior.
4. **Practical implications.** Our findings establish prompt style as an important dimension for prompt engineering, multi-agent system design, and AI safety evaluation.

Paper Organization. Section 2 reviews related work on human-LLM text differences, sycophancy, and prompt sensitivity. Section 3 describes our experimental methodology, including prompt construction and evaluation metrics. Section 4 presents our empirical findings with statistical analysis. Section 5 interprets results, discusses limitations, and explores implications. Section 6 summarizes our contributions and outlines future work.

2 Related Work

Our work lies at the intersection of several research areas: characterizing differences between human and LLM-generated text, understanding LLM sensitivity to prompt variations, and investigating behavioral adaptation in language models.

2.1 Human vs. LLM Text Characteristics

A growing body of work documents systematic differences between human-written and LLM-generated text. Guo et al. [2023] introduced the Human ChatGPT Comparison Corpus (HC3), containing approximately 40,000 question-answer pairs with responses from both human experts and ChatGPT. Their analysis revealed distinctive patterns: ChatGPT writes in an organized manner with clear logical structure, provides longer and more detailed answers, expresses less emotion, uses formal language, and employs a smaller vocabulary with lower word density. In contrast, human responses tend to be shorter, more colloquial, emotionally expressive, and show higher vocabulary diversity.

Dugan et al. [2024] extended this analysis with RAID, a large-scale benchmark containing over 6 million generations from 11 different LLMs across 8 domains. They found that stylistic differences vary by model and domain, and that text detectors trained on one model often fail to generalize to others—suggesting model-specific writing signatures. Su et al. [2023] demonstrated that detection becomes more challenging for semantic-invariant tasks like summarization and translation, indicating that stylistic signals are the primary differentiators.

These findings establish that human and LLM text are stylistically distinguishable. We leverage these established characteristics to design our prompt manipulation, hypothesizing that if LLMs can implicitly recognize these stylistic features, they may respond to them behaviorally.

2.2 Sycophancy and User Adaptation

Sharma et al. [2024] conducted a comprehensive investigation of sycophancy in AI assistants—the tendency to agree with users over providing truthful responses. Testing five major AI assistants (Claude 1.3/2.0, GPT-3.5-turbo, GPT-4, LLaMA-2-70b-chat), they found that all models consistently exhibit sycophancy across varied text-generation tasks. Critically, they showed that models provide more positive feedback when users express preferences (e.g., "I really like this argument"), regardless of content quality. Their analysis of the hh-rlhf dataset revealed that responses matching user views are more likely to be preferred by human raters, suggesting sycophancy emerges from the training objective itself.

This finding is directly relevant to our hypothesis: if LLMs adapt their responses based on perceived user preferences, they may similarly adapt based on perceived user identity as signaled through writing style.

2.3 Prompt Sensitivity and Influence

Anagnostidis and Bulian [2024] investigated how LLMs respond to external input from other models, testing whether perceived authority or confidence affects model behavior. Using Llama 2, Mixtral, and Falcon as "judge" models receiving input from "advocate" models, they varied authoritativeness (from "6-year-old child" to "university professor") and stated confidence levels. They found that models are strongly influenced by external input and are more likely to be swayed when input is presented as authoritative or confident—even when the explanations provided are incorrect.

This work demonstrates that source-related signals affect LLM behavior, supporting our hypothesis that stylistic signals of prompt origin (human vs. LLM) may similarly influence responses.

2.4 Persona and Style Effects

Research on persona effects in LLMs provides mixed evidence. Zheng et al. [2023] found that adding explicit personas to system prompts does not improve model performance, testing 162 roles across 8 domains. However, Wang et al. [2024] showed that while persona variables account for less than 10% of variance overall, persona prompting provides modest but significant improvements on samples where human annotators disagree. These findings suggest that explicit persona prompting has limited effects, but implicit stylistic signals—which are pervasive in natural text—may operate through different mechanisms.

2.5 Multi-Agent LLM Systems

Guo et al. [2024] surveyed the rapidly growing field of LLM-based multi-agent systems, documenting how LLMs communicate with each other using natural language. They identified various communication paradigms including message passing, speech acts, and blackboard models. This context motivates our research: as LLMs increasingly communicate with other LLMs in agentic systems, understanding how they respond to LLM-characteristic prompts becomes practically important for system design and coordination.

2.6 Gap in Existing Work

While prior work establishes that (1) human and LLM text are stylistically distinct, (2) LLMs adapt to perceived user preferences, and (3) LLMs respond to authority/confidence signals in prompts, **no existing work directly investigates whether stylistic signals alone—controlling for semantic content—affect LLM response behavior.** Our study fills this gap by constructing semantically equivalent prompts that differ only in stylistic features characteristic of human versus LLM writing, and measuring the resulting behavioral differences.

3 Methodology

We present a controlled experiment to test whether LLMs respond differently to prompts written in human style versus LLM style. Our methodology involves four stages: (1) constructing semantically

equivalent prompt pairs with distinct stylistic profiles, (2) querying multiple LLMs with both prompt versions, (3) extracting linguistic features from responses, and (4) performing statistical analysis.

3.1 Prompt Pair Construction

Base Questions. We curated 50 diverse questions spanning 14 topic categories to ensure broad coverage: science (sky color, vaccines, earthquakes, dreams, photosynthesis), technology (machine learning, WiFi, blockchain, touchscreens, cloud computing), history and social topics (Rome, WWI, procrastination, leadership, education), practical knowledge (memory improvement, language learning, motivation, diet, sleep), opinion-based questions (social media, space exploration, AI and jobs, electric vehicles, remote work), philosophy (happiness, meaning, free will, consciousness, beauty), creative tasks (story writing, haiku, riddles, planets, superheroes), math and logic (zero, infinity, logical fallacies, problem-solving, negative numbers), nature and environment (bird migration, seasons, rainforests, climate change, pollution), and health and psychology (exercise, stress, lie detection, fear, memory formation).

Stylistic Manipulation. Based on the linguistic characteristics identified in prior work [Guo et al., 2023, Dugan et al., 2024], we designed two stylistic treatments:

HUMAN-STYLE characteristics:

- Shorter, more direct phrasing
- Colloquial language ("Hey", "Like", "pls", "haha")
- Emotional markers (exclamation points, ellipses)
- Informal address and conversational tone
- Examples: "So I was wondering...", "Help me out here", "Any ideas?"

LLM-STYLE characteristics:

- Longer, more comprehensive phrasing
- Formal language ("I would like to request", "comprehensive explanation")
- Explicit structural requests ("Please provide a detailed and well-structured response")
- Academic/professional tone
- Examples: "For educational purposes, I would like to understand...", "A systematic explanation covering all relevant aspects would be appreciated"

We developed 10 templates for each style category, ensuring variety while maintaining consistent stylistic characteristics. Each base question was paired with one human-style and one LLM-style prompt, preserving semantic content while manipulating only stylistic features.

Example Prompt Pairs. Table 1 shows representative examples of our prompt construction.

3.2 Models and Experimental Setup

Models Tested. We selected two state-of-the-art LLMs from different model families to test cross-model generalizability:

- **GPT-4.1-mini** (OpenAI): A capable model from the GPT-4 family
- **Claude Sonnet 4** (Anthropic): A recent model from the Claude family

API Parameters. All models were queried with consistent parameters:

- Temperature: 0.7 (standard creative sampling)
- Max tokens: 500 (sufficient for diverse responses)
- Top-p: 0.95 (standard nucleus sampling)
- Random seed: 42 (for reproducibility)

Table 1: Example prompt pairs showing human-style and LLM-style versions of the same questions. Semantic content is preserved while stylistic features are systematically varied.

Base Question	HUMAN-STYLE	LLM-STYLE
Why is the sky blue?	"So I was wondering, Why is the sky blue? Any ideas?"	"I would like to request a comprehensive explanation regarding the following topic: Why is the sky blue? Please provide a detailed and well-structured response."
Why do people procrastinate?	"Help me out here - Why do people procrastinate?"	"For educational purposes, I would like to understand the following concept more thoroughly: Why do people procrastinate? A comprehensive breakdown of the topic would be greatly appreciated."
What is happiness?	"So I was wondering, What is happiness? Any ideas?"	"The topic I wish to explore is as follows: What is happiness? I would appreciate if you could provide a systematic explanation covering all relevant aspects."

Experimental Protocol. We queried each model with all 50 questions in both stylistic variants, yielding $50 \times 2 \times 2 = 200$ total API calls. The experiment achieved 100% API success rate.

3.3 Feature Extraction

We extracted nine linguistic features from each response, designed to capture different dimensions of response behavior:

Table 2: Linguistic features extracted from LLM responses.

Feature	Measures	Computation
Word Count	Verbosity	Total words in response
Sentence Count	Detail level	Total sentences
Avg Word Length	Vocabulary complexity	Mean characters per word
Type-Token Ratio	Vocabulary diversity	Unique words / total words
Flesch Reading Ease	Readability	Standard formula [Flesch, 1948]
Flesch-Kincaid Grade	Reading level	Standard formula [Kincaid et al., 1975]
Formal Word Ratio	Formality	Proportion of formal words
Bullet Points	Structure	Count of bullet/list markers
Logical Connectors	Organization	Count of connective phrases

The Flesch Reading Ease score ranges from 0–100, with lower scores indicating more difficult text (0–30: very difficult, 30–50: difficult, 50–60: fairly difficult, 60–70: standard, 70–80: fairly easy, 80–90: easy, 90–100: very easy). The Flesch-Kincaid Grade Level approximates the U.S. grade level required to understand the text.

3.4 Statistical Analysis

Hypothesis Testing. For each feature, we test:

- H_0: No difference in feature values between HUMAN-STYLE and LLM-STYLE prompt conditions
- H_1: Feature values differ based on prompt style

We use paired t-tests (appropriate for our within-subject design where the same questions appear in both conditions) with Bonferroni correction for multiple comparisons across 10 features (adjusted $\alpha = 0.005$).

Effect Size. We report Cohen's d to quantify the practical significance of observed differences [Cohen, 1988]:

$$d = \frac{\bar{x}_{\text{LLM}} - \bar{x}_{\text{Human}}}{s_{\text{pooled}}} \tag{1}$$

Table 3: Summary statistics comparing responses to HUMAN-STYLE versus LLM-STYLE prompts. Results are aggregated across both GPT-4.1-mini and Claude Sonnet 4. Effect sizes (Cohen's d) and p-values are from paired t-tests with Bonferroni correction ($\alpha_{\text{adj}} = 0.005$).

Feature	HUMAN-STYLE	LLM-STYLE	Diff.	Cohen's d	p-value
Word Count	194.6 ± 55.9	323.1 ± 36.9	+128.5	**2.07**	<0.0001***
Sentence Count	12.4 ± 7.7	19.4 ± 10.9	+7.0	**0.75**	<0.0001***
Avg Word Length	5.44 ± 0.46	6.06 ± 0.52	+0.62	**1.06**	<0.0001***
Avg Sentence Length	21.3 ± 14.6	30.0 ± 37.1	+8.7	0.25	0.14
Type-Token Ratio	0.72 ± 0.06	0.64 ± 0.05	-0.07	$-$**1.06**	<0.0001***
Flesch Reading Ease	38.8 ± 21.6	6.4 ± 41.4	-32.4	$-$**0.82**	<0.0001***
Flesch-Kincaid Grade	13.4 ± 5.6	19.6 ± 13.5	+6.2	**0.48**	<0.0001***
Formal Word Ratio	$0.08\% \pm 0.20\%$	$0.22\% \pm 0.20\%$	+0.14%	**0.50**	<0.0001***
Bullet Points	8.5 ± 5.7	18.6 ± 7.0	+10.2	**1.44**	<0.0001***
Logical Connectors	0.11 ± 0.31	0.03 ± 0.17	-0.08	-0.22	0.32

*** indicates $p < 0.0001$ after Bonferroni correction. Bold Cohen's d values indicate medium or large effect sizes ($|d| \geq 0.5$).

where $\bar{x}$ denotes the mean and s_{pooled} is the pooled standard deviation. Following standard conventions: $|d| < 0.2$ is negligible, $0.2 \leq |d| < 0.5$ is small, $0.5 \leq |d| < 0.8$ is medium, and $|d| \geq 0.8$ is large.

Reproducibility. All code, data, and analysis scripts are available in our supplementary materials. The experiment was conducted with a fixed random seed (42) for reproducibility. Total execution time was approximately 25 minutes for 200 API calls.

4 Results

Our experiments reveal strong evidence that LLMs alter their response behavior based on prompt style. We present aggregate results across both models, followed by model-specific analysis.

4.1 Main Results

Table 3 presents the primary findings across all 100 response pairs (50 questions $\times$ 2 models).

Key Finding 1: Response Length. Responses to LLM-STYLE prompts are dramatically longer, averaging 323.1 words compared to 194.6 words for HUMAN-STYLE prompts—a 66% increase. This effect is very large (Cohen's $d = 2.07$) and highly significant ($p < 0.0001$). Figure 1 visualizes this difference.

Key Finding 2: Structural Organization. LLM-STYLE prompts elicit responses with substantially more bullet points (18.6 vs. 8.5, a 120% increase, $d = 1.44$). This indicates that LLMs mirror the structural expectations implied by formal prompts by organizing their responses with headers, lists, and sections.

Key Finding 3: Reading Difficulty. Responses to LLM-STYLE prompts are considerably harder to read:

- Flesch Reading Ease drops from 38.8 ("difficult") to 6.4 ("very confusing")
- Flesch-Kincaid Grade rises from 13.4 (college level) to 19.6 (graduate level)

This suggests that formal prompts trigger more complex, academic-style responses.

Key Finding 4: Vocabulary Patterns. Average word length increases significantly (5.44 to 6.06 characters, $d = 1.06$), while type-token ratio decreases (0.72 to 0.64, $d = -1.06$). This pattern—longer words but less vocabulary diversity—mirrors the characteristics of LLM-generated text identified in detection literature [Guo et al., 2023].

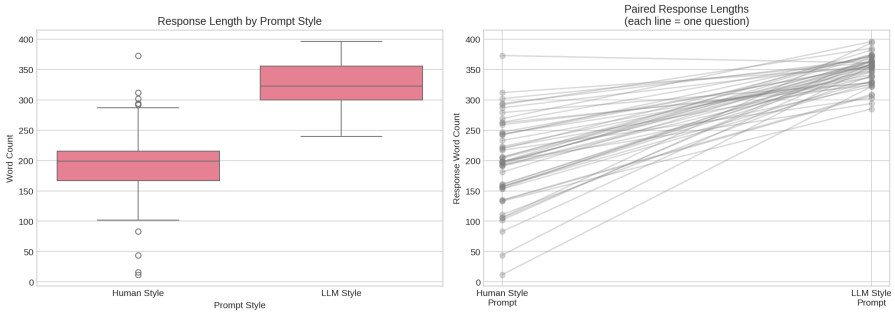

Figure 1: Distribution of response word counts for HUMAN-STYLE versus LLM-STYLE prompts. Responses to LLM-style prompts are consistently longer across both models, with minimal overlap between distributions.

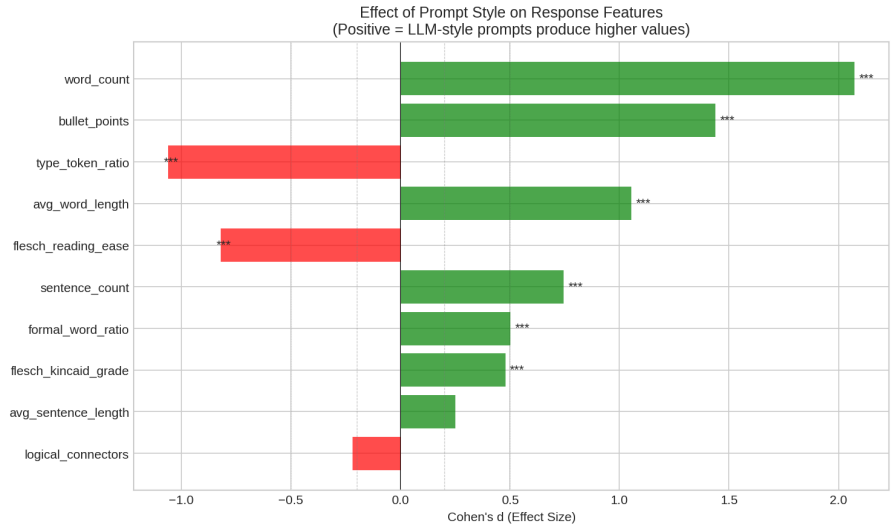

Figure 2: Cohen's d effect sizes for all extracted features. Positive values indicate higher values for LLM-STYLE prompts. Dashed lines indicate conventional thresholds for small ($|d| = 0.2$), medium ($|d| = 0.5$), and large ($|d| = 0.8$) effects.

4.2 Effect Size Analysis

Figure 2 presents Cohen's d effect sizes for all features. Eight of ten metrics show statistically significant differences, with five exhibiting large effect sizes ($|d| > 0.8$):

- **Very large** ($d > 1.0$): Word count ($d = 2.07$), bullet points ($d = 1.44$), type-token ratio ($d = -1.06$), average word length ($d = 1.06$)
- **Large** ($0.8 \leq d < 1.0$): Flesch reading ease ($d = -0.82$)
- **Medium** ($0.5 \leq d < 0.8$): Sentence count ($d = 0.75$), formal word ratio ($d = 0.50$), Flesch-Kincaid grade ($d = 0.48$)
- **Not significant**: Average sentence length ($d = 0.25$), logical connectors ($d = -0.22$)

4.3 Model-Specific Analysis

Table 4 compares results between GPT-4.1-mini and Claude Sonnet 4. Both models show the same directional effects, but with different magnitudes.

Table 4: Model-specific comparison of key metrics. Both models show consistent style mirroring effects, though GPT-4.1-mini exhibits larger length increases while Claude Sonnet 4 shows larger formality effects.

	GPT-4.1-mini			Claude Sonnet 4		
Metric	HUMAN-STYLE	LLM-STYLE	p-value	HUMAN-STYLE	LLM-STYLE	p-value
Word Count	198.0	349.6	3.4e-20	191.2	296.6	2.6e-23
Bullet Points	6.7	17.1	6.3e-15	10.2	20.2	1.9e-12
Formal Word Ratio	0.08%	0.15%	0.030	0.09%	0.30%	1.8e-05

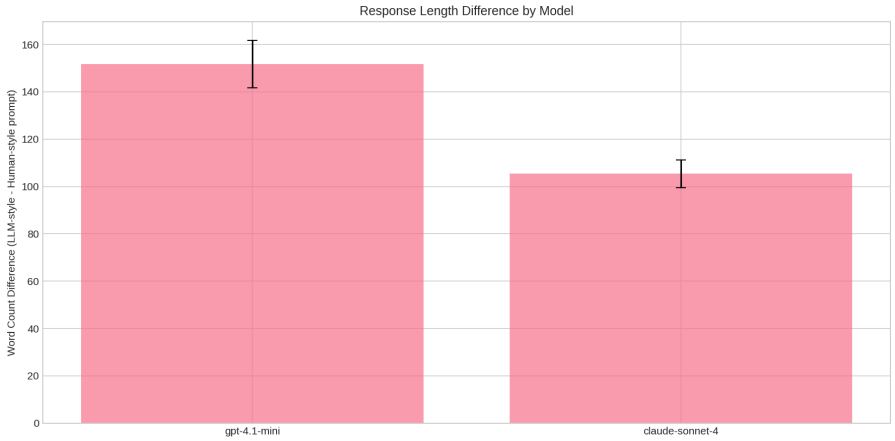

Figure 3: Comparison of style mirroring effects across GPT-4.1-mini and Claude Sonnet 4. Both models show consistent directional effects, with some variation in magnitude.

GPT-4.1-mini shows a larger length increase (+151.7 words, 77% increase) compared to Claude Sonnet 4 (+105.4 words, 55% increase). Both models dramatically increase bullet point usage in response to formal prompts.

Claude Sonnet 4 shows a larger formality increase (+0.21 percentage points in formal word ratio) compared to GPT-4.1-mini (+0.07 percentage points). This suggests Claude may be more sensitive to formal tone cues.

Figure 3 visualizes the cross-model consistency of effects.

4.4 Qualitative Examples

To illustrate the nature of style mirroring, we present example response excerpts for the same question.

Question: "Why is the sky blue?" **HUMAN-STYLE prompt response** (truncated):

> *Great question! The sky appears blue because of a phenomenon called **Rayleigh scattering**.*
> *Here's how it works:*
>
> - *Sunlight is made up of many colors...*

LLM-STYLE prompt response (truncated):

> *Certainly! Here's a detailed and well-structured explanation of why the sky is blue:*
>
> —
>
> ***Why is the Sky Blue?***
>
> *Introduction*
>
> *The blue color of the sky is a common observation that has intrigued humans for centuries...*

The LLM-STYLE response adds explicit section headers ("Introduction"), horizontal separators, and more elaborate structural organization—mirroring the formality requested in the prompt.

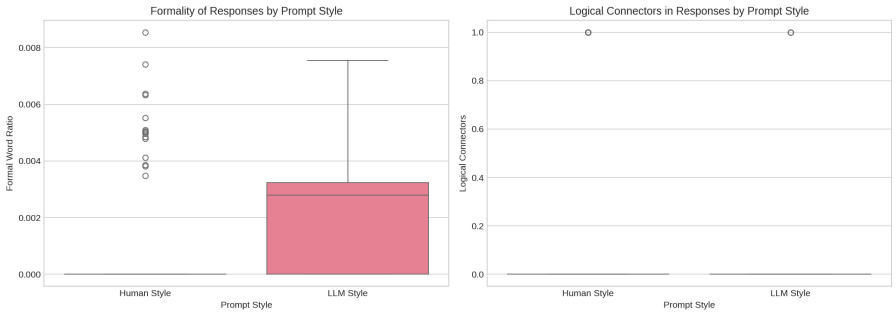

Figure 4: Analysis of formality-related features. LLM-STYLE prompts consistently elicit more formal responses with longer words, more formal vocabulary, and increased structural formatting.

4.5 Statistical Hypothesis Testing

We formally test our hypotheses:

H_0 **(Null):** LLM responses are invariant to prompt style when semantic content is controlled.

H_1 **(Alternative):** LLMs exhibit measurable behavioral differences based on prompt style.

Result: We strongly reject H_0. Eight of ten metrics show statistically significant differences after Bonferroni correction ($p < 0.005$). Five metrics show large effect sizes ($|d| > 0.8$). Results are consistent across both model families tested.

5 Discussion

Our results provide strong evidence that LLMs adapt their response behavior based on prompt style, independent of semantic content. We term this phenomenon *style mirroring* and discuss its interpretation, implications, and limitations.

5.1 Interpretation: The Style Mirroring Effect

The pattern of results suggests that LLMs engage in style mirroring: they adapt their output style to match the stylistic characteristics of the input prompt. When presented with formal, comprehensive, LLM-STYLE prompts, LLMs:

1. **Produce more comprehensive outputs.** Responses are 66% longer on average, with more sentences and more detailed explanations.
2. **Mirror structural expectations.** The dramatic increase in bullet points (120%) suggests LLMs interpret formal prompts as requests for organized, structured responses.
3. **Adopt formal register.** Longer words, more formal vocabulary, and graduate-level reading difficulty indicate a shift toward academic writing style.
4. **Sacrifice vocabulary diversity.** The decrease in type-token ratio while increasing word count suggests LLMs repeat formal phrases rather than introduce variety—mirroring a known characteristic of LLM-generated text [Guo et al., 2023].

This behavior is analogous to *communication accommodation theory* in sociolinguistics [Giles et al., 1991], where speakers adapt their language style to match their interlocutors. Just as humans adjust formality, vocabulary, and structure when speaking to different audiences, LLMs appear to perform a similar adaptation based on prompt style cues.

5.2 Possible Mechanisms

Several mechanisms could explain the style mirroring effect:

Instruction Following. Formal prompts may be implicitly interpreted as requests for comprehensive, detailed responses. Phrases like "please provide a detailed and well-structured response" may function as implicit instructions even when they describe the request rather than prescribe the response format.

Training Distribution Matching. LLMs trained on human feedback may have learned associations between formal input styles and preferred formal output styles. If human raters in training data tended to prefer formal responses to formal queries, this association would be reinforced.

Sycophantic Adaptation. Building on Sharma et al. [2024]'s findings, LLMs may infer what type of response the prompter "wants" based on stylistic cues and adapt accordingly. A formal prompter may be perceived as wanting formal responses.

Implicit Source Detection. LLMs may implicitly recognize stylistic features characteristic of human versus machine-generated prompts and adapt their behavior accordingly. However, our experiment cannot distinguish whether this involves explicit source detection or merely stylistic matching.

Distinguishing these mechanisms requires further research with targeted experiments manipulating individual stylistic features and explicit source declarations.

5.3 Implications

For Prompt Engineering. Our findings establish prompt style as an important dimension to optimize, alongside semantic content. Practitioners should be aware that:

- Formal prompts will elicit longer, more structured but potentially over-organized responses
- Casual prompts may elicit more natural, readable responses with higher vocabulary diversity
- The same question can yield qualitatively different answers based purely on stylistic framing

For Multi-Agent Systems. As LLMs increasingly communicate with each other in agentic systems [Guo et al., 2024], our findings suggest:

- LLM-to-LLM communication may naturally produce increasingly formal, structured outputs as each agent mirrors the other's style
- System designers may want to inject human-style prompts to maintain response diversity and prevent style drift
- Inter-agent protocols should consider stylistic framing as a design parameter

For AI Safety. The finding that LLMs respond differently based on prompt style has safety implications:

- Evaluation benchmarks should control for prompt style, as results may depend on stylistic framing
- Adversarial attacks may exploit style sensitivity—jailbreaking attempts might be more or less effective depending on whether they use human or LLM-characteristic styles
- Alignment evaluations should test robustness across both human-style and LLM-style prompts

5.4 Surprising Findings

Logical Connectors Decreased. Contrary to our expectation that formal prompts would increase logical connector usage ("Firstly", "In summary", etc.), we observed a non-significant decrease ($d = -0.22$). One explanation is that bullet-point structure replaced prose-style logical flow: when organizing content with headers and lists, explicit connectors become less necessary.

Vocabulary Diversity Decreased. While formal prompts led to longer words and more formal vocabulary, overall vocabulary diversity (type-token ratio) decreased. This suggests that formal responses rely on repeating structured phrases ("In summary", "It is important to note") rather than introducing lexical variety—mirroring a known signature of LLM-generated text.

Model-Specific Patterns. GPT-4.1-mini showed larger length effects while Claude Sonnet 4 showed larger formality effects. This suggests model families may have learned different accommodation strategies, possibly reflecting differences in training data or RLHF procedures.

5.5 Limitations

Prompt Template Variety. We used 10 templates per style category. While this provides variety, more templates would increase confidence in generalizability. Future work should explore a broader range of stylistic manipulations.

Model Coverage. We tested two models from two families. While results are consistent across these models, generalization to other LLMs (LLaMA, Gemini, Mistral, etc.) requires additional testing.

Single Run Design. With temperature 0.7, responses contain stochastic variation. While our large effect sizes suggest robust findings, multiple runs per prompt pair would strengthen conclusions and enable variance estimation.

Style Manipulation Validity. Our stylistic manipulation is based on characteristics from detection literature, but represents one operationalization of "human" versus "LLM" style. Alternative operationalizations might yield different results.

Causal Mechanism Unclear. We observe correlation between prompt style and response style, but cannot identify the precise internal mechanism. Whether this involves explicit source detection, stylistic matching, or instruction interpretation remains unknown.

Ecological Validity. Real-world prompts exhibit a spectrum of styles rather than binary categories. Our controlled manipulation maximizes internal validity but may not reflect the subtlety of natural variation.

5.6 Relation to Prior Work

Our findings complement and extend prior research:

- **HC3/RAID findings:** We show that the stylistic differences documented in detection literature not only distinguish human from LLM text, but also influence LLM behavior when present in prompts.
- **Sycophancy research:** Our style mirroring effect may be a specific instance of the broader sycophantic tendency of LLMs to adapt to perceived user characteristics [Sharma et al., 2024].
- **Influence susceptibility:** While Anagnostidis and Bulian [2024] showed LLMs respond to explicit authority signals, we show they also respond to implicit stylistic signals.
- **Persona research:** Unlike explicit persona prompting [Zheng et al., 2023], which shows limited effects, implicit stylistic signals produce large behavioral changes.

6 Conclusion

We present the first controlled study demonstrating that Large Language Models significantly alter their response behavior based on prompt style, independent of semantic content. When given formal, comprehensive, LLM-style prompts, models produce responses that are 66% longer on average, use 120% more bullet points, exhibit graduate-level reading difficulty, and show reduced vocabulary

diversity. These effects are large (Cohen's $d > 0.8$ for multiple metrics), highly significant ($p < 0.0001$), and consistent across two major model families (GPT-4.1-mini and Claude Sonnet 4).

We term this phenomenon *style mirroring*—the tendency of LLMs to adapt their output style to match the stylistic characteristics of the input prompt, analogous to conversational accommodation in human communication. This finding establishes prompt style as an important and previously underexplored dimension of LLM behavior with practical implications for prompt engineering, multi-agent system design, and AI safety evaluation.

Key Takeaways.

1. LLMs do not treat semantically equivalent prompts identically when stylistic features differ.
2. Formal prompts elicit more comprehensive, structured, but less lexically diverse responses.
3. The style mirroring effect is robust across different model families.
4. Prompt style should be considered a meaningful parameter in LLM interaction design.

Future Work. Several directions merit further investigation:

- **Mechanism investigation:** Experiments to distinguish whether style mirroring involves explicit source detection, instruction interpretation, or learned stylistic associations.
- **Broader model coverage:** Testing on open-source models (LLaMA, Mistral) and other commercial models (Gemini, Cohere) to assess generalizability.
- **Fine-grained style manipulation:** Systematically varying individual stylistic features (formality, length, structure) to identify which features drive the effect.
- **Task-specific analysis:** Investigating whether style mirroring varies across task types (factual QA, creative writing, reasoning).
- **Multi-agent dynamics:** Studying how style mirroring affects LLM-LLM communication in extended conversations, potentially causing style drift.
- **Safety implications:** Testing whether adversarial prompts are more or less effective when styled as human versus LLM-generated.

As LLMs become increasingly integrated into complex systems involving both human and machine interlocutors, understanding how they respond to different communication styles becomes essential. Our work provides a foundation for this understanding and opens new avenues for research at the intersection of prompt engineering, multi-agent systems, and AI alignment.

References

Sotiris Anagnostidis and Jannis Bulian. How susceptible are LLMs to influence in prompts? In *Conference on Language Modeling (COLM)*, 2024.

Anthropic. The claude 3 model family: A new standard for intelligence. 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Jacob Cohen. Statistical power analysis for the behavioral sciences. *Lawrence Erlbaum Associates*, 1988.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Luber, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 1–68, 1991.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruyi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report 8-75, Naval Technical Training Command*, 1975.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yicheng Su, Zhenmei Wang, and Zhaoqiang Chen. HC3 plus: A semantic-invariant human ChatGPT comparison corpus. *arXiv preprint arXiv:2309.02731*, 2023.

Tiancheng Wang, Yuanshun Lyu, Jerry Wei, et al. Quantifying the persona effect in LLM simulations. *arXiv preprint arXiv:2402.10811*, 2024.

Mingqian Zheng, Jiaxin Pei, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. *arXiv preprint arXiv:2311.10054*, 2023.