Step-By-Step Reasoning with Meta Cognitive Prompts to Reduce Contextual Hallucination

Brian Miki
Simon Fraser University
School of Computing Science
Canada
brian_miki@sfu.ca

Nicholas Vincent
Simon Fraser University
School of Computing Science
Canada
nvincent@sfu.ca

Abstract

The advancement in Large Language Models (LLMs) and downstream AI products have significantly improved automation use cases for both individual consumers and large enterprises. However, a significant challenge exists in the form of contextual hallucinations: when a model generates an output that is contextually incorrect and irrelevant to the input. This paper focuses on developing a reasoning strategy that guides the LLM through a series of cognitive steps to reduce contextual hallucination as well as measuring the impact of the strategy on LLM performance in multi-step reasoning tasks in the domain of outreach emails. To evaluate LLM performance, we curated a dataset of outreach email tasks that required reasoning across multiple steps including logical structuring, personalization and recipient alignment. We provide early evidence that structured cognitive prompts improve agreement between human evaluators on LLM generated outreach emails, demonstrating its effectiveness in reducing contextual hallucinations. We highlight the critical role of ongoing human evaluation and practices from HCI for integrating LLMs into enterprise workflows.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; • **Computing methodologies** → *Natural language processing*; *Evaluation methodologies.*

Keywords

Large Language Models, Prompt Engineering, System 1 and System 2 Thinking, Metacognition, Outreach Emails, Contextual Hallucination, Human Evaluation

ACM Reference Format:
Brian Miki and Nicholas Vincent. 2025. Step-By-Step Reasoning with Meta Cognitive Prompts to Reduce Contextual Hallucination. In *HEAL @ CHI '25: Proceedings of the Human-centered Evaluation and Auditing of Language Models, April 26, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

HEAL @ CHI '25, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/25/04
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

LLMs continue to advance and have seen more integration more heavily into software systems. On the capability side, LLMs have seen improvements in synthesizing numerical data, summarizing information which require handling various inputs and question-answering. In the enterprise space, these advancements have opened up the possibilities of AI workflows which help companies to automate tasks that vary in difficulty from simple one sentence replies to complete engagements with customers. In the consumer space, AI has improved capabilities in virtual assistants and chatbots which have the ability to perform tasks such as setting reminders, answer questions and managing smart home devices[4].

While extremely impressive in performance, LLMs are still without certain constraints. A significant challenge for LLMs is their ability to provide concise answers that align contextually with the input of the user [11]. Contextual hallucinations in this context refers to a LLM generating a response that is out of context or irrelevant to the input prompt. For software systems such as virtual assistants and AI workflows to be production ready, contextual hallucination must be mitigated or removed completely[6].

Reducing the risk of contextual hallucination is crucial for improving the reliability and accuracy of using LLMs and their related accuracy. Contextual hallucinations can lead to a lack of trust, unfocused results and misinformation. These issues lead to a lack of adoption in large enterprise AI systems and consumer products [5].

Many techniques have been researched to reduce the frequency of contextual hallucinations such as fine-tuning with catered data sets, retrieval augmented generation (RAG) and prompt engineering. However, these solutions require a large amount of manual effort, capital and time. Thus, a low cost and widely adoptable list of best practices is required to advance research on mitigating contextual hallucinations and increasing adoption of LLMs and their AI systems. What our research focuses on is developing a multi-step reasoning process for LLMs to improve reasoning and reduce contextual hallucination.

To systematically evaluate the impact of contextual hallucination and explore strategies for improvement, we curated a dataset of outreach email generation tasks that simulate real-world engagements. The dataset consists of various outreach situations, including sales outreach, where the goal is to engage potential clients with a product and value proposition; venture capital fundraising, which involves presenting a pitch to an investor using business metrics; and student networking emails, which focus on connecting with professionals for mentorship opportunities.

For researchers interested in LLMs for productivity – for instance, improving outreach emails – this evaluation approach suggests

some actionable insights regarding when to prompt for System 1 and System 2 thinking. Below, we further describe key work that inspired our approach, our methods, the human-centered evaluation process, and our results. Finally, we discuss how contextual hallucinations might be reduced, and other implications of LLMs for productivity.

2 Related Studies

2.1 Prompt Engineering and Related Techniques

Prompt Engineering and the design of a prompt and its inputs can drastically impact response output. When using a general prompt with little information the LLM lacks context resulting in a generic response. Providing instructions, context and inputs allows for more specific and accurate outputs which reduce the likelihood of contextual hallucination. Creating prompts that are clear and precise decrease ambiguity and guide the LLM toward a more desired generated output [15].

Role-prompting is another technique fundamental to prompt engineering and reducing contextual hallucination. This involves giving the model a specific role to act as such as a customer service representative or library assistant. This method provides context to the role and knowledge the LLM will utilize, ensuring an output that aligns more effectively with the desired output [9]. For example, if the model is prompted to act as a long-time financial advisor, the LLM is more likely to provide a more detailed and contextually accurate response to topics around financial advice and learnings from previous economics.

Prompt chaining is the process of breaking down a set of instructions for a prompt into separate outputs. Allowing the LLM to abstract certain steps into its individual output can allow for better reasoning, more detailed responses and an increase in accuracy. Prompt chaining is effective in complex problem solving tasks, ambiguous prompts and multi step processes as it breaks down problems into various steps. Through a clearly defined process, prompt chaining enables for more transparency in the reasoning and process allowing for simpler iteration processes [12] Thought propagation is a technique inspired by the analogical reasoning process, where LLMs follow a stepwise approach to complex reasoning tasks by breaking down the problem into separate tasks and utilizing the processes to solve each to complete a more complex task [13].

2.2 Cognitive Psychology & Bias in LLM Prompting

Inspired by human cognitive psychology, System 1 and System 2 thinking refers to the distinct modes of reasoning LLMs can be prompted to simulate. System 1, fast thinking, deals with intuitive responses, found in straightforward and well-known tasks. In contrast, System 2, or slow thinking, requires more analytical thought for complex scenarios [14].

Beyond hallucinations, LLMs also express patterns similar to human cognitive biases[10]. These biases influence reasoning and decision-making in LLM outputs. Cognitive biases refer to systematic deviations from rational judgment, and their presence in LLMs introduce inconsistencies in responses[2]. Similar to how contextual hallucination reduces LLM reliability, cognitive biases can lead to inconsistent or inaccurate responses in outreach emails, where structured reasoning and personalization are essential.

3 Methodology

To assess the reasoning capabilities of LLMs, we curated a dataset of tasks specifically focused on generating outreach emails. These tasks were designed to require elements of both System 1 thinking-providing an empathetic, human aspect-and System 2 thinking-ensuring logical structure and coherence-creating a robust dataset for evaluating LLM performance and its reasoning capabilities.

We designed a two iteration experiment using a fixed set of outreach email generation tasks across three scenarios: sales outreach, venture capital fundraising, and student networking. In both iterations, the underlying prompt structure and profile inputs remained constant.

In the first iteration, the LLM was given a general instruction to generate a professional outreach email using the structured profile data, with no specific guidance on reasoning or tone.

In the second iteration, we introduced a meta cognitive prompting strategy that explicitly invoked System 2 and System 1 thinking in sequence. The prompt was structured to first guide the LLM through a reflective, analytical process (System 2), prompting it to consider the recipient's perspective, values, desired tone, appropriate length, and logical structure of the email. This encouraged deliberate planning before writing. The prompt then shifted to invoke System 1 thinking, instructing the LLM to apply an intuitive and empathetic tone when writing the final message.

Two human-raters evaluated LLM performance across the three outreach scenarios: sales outreach, venture capital fundraising and student networking emails. Across these three scenarios, we assessed a total of 180 LLM generated outreach emails, 60 per scenario analyzing improvements in structured reasoning, personalization, and alignment with recipient expectations.

To evaluate the effectiveness of LLM-generated outreach emails, we curated a dataset that reflects Western business culture and communication norms. The dataset includes detailed sender and recipient profiles designed to resemble common business interactions, roles, and processes found in Western corporate environments.

Our curation process was influenced by how professionals in these settings typically engage in outreach - focusing on aspects such as direct and personable communication, value-driven propositions, and industry-specific pain points. The information included in each profile is based on details that would be readily available in professional interactions, such as LinkedIn profiles, and company websites. Each profile was artificially created but informed by real-world data sources

By structuring the dataset this way, we ensure that the generated outreach emails reflect realistic, contextually appropriate messaging strategies that resonate with Western professionals. This approach allows us to assess how well LLMs adapt to industry norms, personalize outreach, and align with expectations in Western business communication.

Table 1: Sales Outreach Details

Sender Name	Jessica Martin
Sender Role	Sales Executive
Sender Background	CloudSync expert in cloud storage solutions.
Recipient Name	Mark Sullivan
Recipient Role	IT Director
Recipient Company	Zenith Manufacturing
Industry	Manufacturing
Objective	Schedule a 30-minute discovery call.
Value Proposition	CloudSync improves document workflow, reducing management time by 40%.
Personal Connection	Shared interest in optimizing team collaboration.
Professional Connection	Jessica specializes in cloud storage; Mark manages IT.
Case Study	Assisted a manufacturer in cutting file management times by 40%.
Recent Achievement	Launched an AI-powered document search tool.
Recipient's Goal Alignment	CloudSync enhances productivity and collaboration.
Call to Action	Propose a 30-minute call.

Each entry contained fields such as sender background, recipient role, professional alignment, value proposition, and call-to-action objectives. Highlighted in table 1 is an example of the structured data used for email generation and evaluation:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

To objectively assess the effectiveness of Large Language Model (LLM) outputs in generating outreach emails, we employ human evaluation of quality, and use Cohen's Kappa Coefficient to measure agreement. This metric is utilized in our research to measure the agreement between two human evaluators who looked at all the outputs produced, providing a robust assessment of the LLM's performance relative to human expectations and criteria we set forth. Our approach aligns with recent advancements in LLM evaluation frameworks[8].

To calculate Cohen's Kappa Coefficient, two key values are measured. Observed agreement is the proportion of cases where both evaluators assign the same rating to an email. Expected agreement measures the agreement that would occur by chance alone if the two evaluators rated emails independently based on their own tendencies to choose ratings. Using two human evaluators, some level of agreement will occur naturally by chance, the expected agreement accounts for this ensuring that the Cohen's Kappa Coefficient measures true agreement rather than coincidence.

Cohen's Kappa Coefficient ranges from -1 to +1 with +1 indicating perfect agreement that the LLM output aligns with human expectations for effective outreach emails. -1 signifies complete disagreement, indicating the output is not at all aligned with human expectations and for an effective outreach email.

The overall evaluation process is defined in two core segments. We assess structural components including the quality of specific sections. These components are deemed essential for crafting an effective outreach email. The evaluation criteria includes: subject line effectiveness, opening paragraph engagement, body content relevance and call to action clarity.

Additionally we evaluate qualitative aspects of the LLM outputs. These aspects are more subjective but contribute to the LLMs overall effectiveness. The evaluation criteria includes: tone appropriateness, personalization and conciseness.

For our evaluation process: two independent human evaluators assess each LLM-generated outreach email using a standardized rubric based on the above criteria. Their assessments are then analyzed using Cohen's Kappa to determine the level of agreement and, by extension, the effectiveness of the LLM output.

This dual approach allows for a comprehensive evaluation of the outreach emails, providing a holistic view of the LLMs performance in this task. Additionally, the distribution of ratings is measured to calculate improvements in specific categories of evaluation when going from one iteration to another.

4 Results

4.1 Cohen's Kappa Coefficient Results

In addition to consistency, substantial improvements were noted in specific email evaluation categories, particularly in scenarios where the metacognitive prompting strategies were applied.

System 1 Thinking: Prompts that used intuitive outputs significantly improved the tone and engagement such as LLM-generated sales outreach emails. Evaluators noted more relatable, conversational tones, which contributed to improved subject lines and call-to-action effectiveness. The strong agreement (Kappa > 0.5) in these cases highlights the effectiveness of these changes.

System 2 Thinking: Structured prompts that guided the model through logical steps led to notable improvements in categories like Structure, Concision, and Call to Action clarity. For complex emails, such as those targeting venture capital funding, logical structuring provided clarity and coherence. However, improvements in conciseness were less noticeable, with minimal reduction in character count.

4.2 Interpreting Disagreement in Evaluation

While Cohen's Kappa provides a useful quantitative measure of agreement, the instances of disagreement observed between raters are also important to highlight. When evaluating LLM-generated outreach emails, disagreement does not necessarily indicate poor evaluation quality.

Outreach emails blend factual, persuasive, and emotional content. When a model generates a message, evaluators may legitimately differ in how they assess aspects such as tone, conciseness, or relevance. These discrepancies highlight the variability in how recipients might interpret messages in the real world, an important consideration for LLM adoption and human centered evaluation.

Table 4: Inter-Rater Disagreement on Concision Across 15 VC Outreach Email Generations

Email	Rater A	Rater B
1	$4 \rightarrow 3$ ($\downarrow$)	$3 \rightarrow 4$ ($\uparrow$)
2	$5 \rightarrow 3$ ($\downarrow$)	$4 \rightarrow 5$ ($\uparrow$)
3	$3 \rightarrow 3$ (–)	$2 \rightarrow 3$ ($\uparrow$)

Table 2: Improvement in Rater Evaluation Per Task Set (%)

Category	Subject Line	Introduction/Hook	Your Background	Value Highlighted	Call to Action	Structure	Concision	Tone
Sales	26.01	43.20	10.39	1.85	29.91	24.33	26.00	40.12
Student	2.22	12.58	5.54	6.67	10.77	5.97	12.47	13.94
VC	12.21	42.61	49.81	5.76	3.85	13.34	-4.57	20.59

Table 3: Cohen's Kappa Coefficient Calculation Per Task Set

Scenario	Observed Agreement (P_o)	Expected Agreement (P_e)	Cohen's Kappa (κ)	Interpretation
Sales Outreach	0.625	0.639	-0.038	Slight disagreement
Iteration 2	0.875	0.750	0.500	Moderate agreement
Student Networking	0.833	0.611	0.571	Moderate agreement
Iteration 2	0.833	0.486	0.676	Strong agreement
VC Outreach	0.750	0.611	0.357	Fair agreement
Iteration 2	0.833	0.510	0.660	Moderate to strong agreement

One case of inter-rater disagreement emerged in the venture capital outreach scenario. Here, the conciseness category saw a negative shift of -4.57%, the only decline across all evaluation criteria. One evaluator valued the additional business metrics and structured pitch framing, viewing it as appropriate detail for a high-context audience such as investors. However, the other rater interpreted these details as unnecessary. This disagreement highlights how even experienced evaluators may interpret information differently depending on context, audience expectations, and personal communication norms.

Rather than viewing disagreement as a failure in improvement, we interpret it as evidence of where human interpretation can differ, especially in tasks requiring additional context. This aligns with the goals of human-centered AI evaluation, understanding not only what the model outputs but how people understand it.

5 Discussion

Our findings reveal improvements in the application of System 1 and System 2 thinking, which impact the quality of LLM-generated outreach emails through step by step reasoning. Here, we discuss some of the implications of these findings and this line of work, including (1) applications areas where experimenting with prompting may be useful and the trade-offs involved with different styles of prompts. Then we focus on limitations of our approach, including the general limitations of using the System 1/2 framework (both in terms of its generalization to people and its connection to LLM mechanics), potential interactions with cultural differences in business practices, and the impact of the rapidly-changing landscape of AI products on AI prompting and evaluation practices.

In terms of System 1, we observed large improvements in emails requiring a personal, empathetic touch, such as those used in sales outreach. By guiding the LLM to simulate fast, intuitive reasoning, emails became noticeably more relatable and engaging, with a tone that resonated more closely with a human conversational style. However, this improvement was less noticeable in more transactional types of outreach emails, such as venture capital (VC) funding requests. These communication types may rely less on emotional appeal and more on structured, factual information, which reduces the relevance of System 1 thinking.

For System 2 thinking, which involves a deeper level of logic and analytical reasoning, improvements were less evident. As the value proposition was often clearly stated, we saw slight enhancements in the model's ability to adjust the value propositions and cater to the recipient.

In addition to these qualitative differences, we also examined the structural quality of the outputs. While System 1 thinking contributed to more engaging and authentic sales pitches, improvements in conciseness were limited across all types of outreach emails. Despite instructing the LLM to prioritize conciseness, minimal changes in character count were observed. This limitation may stem from placing the conciseness prompt at the end of the reasoning structure or may indicate that LLMs require more granular guidance on improving concision.

A consideration to mention in implementing a reasoning process that implements System 2 and System 1 thinking is the increased token usage associated with more complex, multi-step reasoning tasks. By prompting the model to follow a deliberate and analytical process we inevitably increase the total token count which impacts both processing time and cost. While this cost may be justified for high-value interactions where precision and clarity are critical it may not be as beneficial for simpler tasks where a single-step response is appropriate.

Utilizing System 2 thinking presents a trade-off: the minimal gains in logical coherence and relevance must be weighed against the higher token expenditure. For many enterprise applications, particularly those involving repetitive or straightforward inquiries, the increased token cost may not justify the benefits. However, for tasks where depth, precision, and contextual relevance are key, the added investment in System 2 thinking can lead to significantly improved outcomes and a more sophisticated interaction experience.

Breaking down interactions into System 1 and System 2 thinking could provide a useful framework for structuring outreach, but it also comes with some risk of oversimplification and potentially misleadings users about the actual mechanistic operation of LLM tools. Communication is complex as it is shaped by cultural norms, creativity, and adaptability. In many cases, outreach efforts require a combination of both intuitive and analytical thinking, making it

difficult to categorize every aspect of an interaction into one system or the other.

5.1 Limitations

System 1 and System 2 thinking holds the risk of oversimplifying human cognition and misleading users in the mechanistic operation of LLMs and their outputs. Systems thinking provides a useful example for understanding human cognition, however, oversimplifies how humans reason. Human reasoning does not operate in two clearly defined modes but rather exists where intuition and deep reasoning work together in tandem rather then sequentially ([7]. Similarly, for LLMs, applying System 1 and System 2 thinking is potentially misleading as LLMs do not have separate processes and operations for fast intuition and slow deliberation. While their ability to mimic human performance in outreach emails can be seen as high, there is an illusion of reasoning and no shift between the two cognitive systems. Misinterpreting LLM outputs through a System 1/2 lens may lead users to misunderstand LLM mechanistic operations as deliberative thought instead of probabilistic token output.

One of the key limitations of our study is that it was designed based on Western outreach norms. Scholars have claimed that In Western culture direct communication, clear value propositions, and structured calls to action are prioritized. However, Eastern business cultures, such as those in Japan, operate differently, meaning that a System 1 and System 2 approach may not be as effective [3].

Japanese business culture places a strong emphasis on hierarchy, relationship building, and indirect communication. Decision making is often consensus driven, and transactional outreach can be seen as overly aggressive [3]. If an LLM generated outreach email follows a rigid structure that assumes a Western approach, it may fail to resonate with a Japanese email recipient. This cultural misalignment highlights one of the core limitations of applying System 1 and System 2 thinking universally, as different cultural contexts may require different communication strategies.

Another major challenge in structuring outreach using System 1 and System 2 thinking is that it does not always include unconventional engagement strategies. Many successful outreach efforts rely on storytelling or humor which do not fit perfectly into intuitive or analytical thinking. Humor, in particular, is difficult to categorize into a structured cognitive process because it depends on timing, tone, and social context. If an LLM is trained to follow a rigid System 1 and System 2 structure, it may fail to generate outreach messages that feel spontaneous or genuinely engaging.

As the AI landscape continues to evolve, the rise of new models like DeepSeek [1] highlight how advancements in architecture, reasoning capabilities, and contextual understanding reshape the way we approach prompting strategies and human centered evaluation methodologies. While the System 1 and System 2 framework has proven effective for structuring outreach emails with existing models, we must remain cautious about assuming its long-term applicability. As newer models become more capable at handling complex reasoning and nuanced communication.

Given this rapid progression, it is crucial to adopt a flexible and iterative approach to prompting - one that progresses with emerging AI capabilities. As models continue to change, so must our strategies for utilizing them effectively. We should focus on continuously testing, refining, and adapting our approaches to align with the latest advancements in AI reasoning, human evaluation and contextual understanding.

6 Conclusion

Our research demonstrates that guiding Large Language Models (LLMs) through structured System 1 and System 2 thinking can impact their effectiveness in generating engaging responses that reduce contextual hallucination. While System 1 thinking lends itself to tasks requiring empathy and a personal touch, System 2 thinking supports logical coherence and analytical reasoning. By employing these strategies in a structured, step-by-step manner, we can potentially reduce contextual hallucinations and create outputs that better align with user expectations personally and professionally.

Overall, our findings highlight the value of a cognitive, metacognitive prompting strategy that could enhance user trust and satisfaction and reduce contextual hallucination.

Acknowledgments

The authors would like to thank the Simon Fraser University and the School of Computing Science for the resources provided.

References

[1] DeepSeek-AI et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*. https://arxiv.org/abs/2501.12948
[2] Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. University of California, San Diego & MIT-IBM Watson AI Lab. *arXiv preprint arXiv:2403.00811*.
[3] Hooker, J. (2008). Cultural differences in business communication. Carnegie Mellon University. https://public.tepper.cmu.edu/jnh/businessCommunicationDifferences.pdf
[4] Huang, S. H. (2023, July 19). Effectively harnessing AI and workflow automation to improve the customer experience. Forbes. https://www.forbes.com/sites/forbestechcouncil/2023/07/19/effectively-harnessing-ai-and-workflow-automation-to-improve-the-customer-experience/
[5] Huang, L., Jiang, Y., Xie, N., Liu, Q., Lu, Y., Huang, X., & Gao, J. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2403.05232*.
[6] Jo, E., Kim, Y.-H., Jeong, Y., Park, S., & Epstein, D. A. (2025). Incorporating multi-stakeholder perspectives in evaluating and auditing of health chatbots driven by large language models. *Proceedings of the HEAL Workshop at CHI 2025*.
[7] Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
[8] Kim, T. S., Lee, Y., Shin, J., Kim, Y. H., & Kim, J. (2024). EvalLM: Interactive evaluation of large language model prompts on user-defined criteria. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642216
[9] Li, Z., Yu, W., Wang, Y., & Ren, X. (2023). Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.
[10] Oza, J., & Yadav, H. (2024). Metrics to meaning: Enabling human-interpretable language model assessment. *Proceedings of the HEAL Workshop*.
[11] Ravuri, S., d'Autume, C. D., Altinok, A., & von Wichert, F. (2024). Language model pre-training at scale: Unveiling the secrets of context, structure, and generalization. *arXiv preprint arXiv:2403.09704*.
[12] Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., & Cai, C. J. (2022). PromptChainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–10). ACM.
[13] Yu, J., He, R., & Ying, R. (2023). Thought propagation: An analogical approach to complex reasoning with large language models. *arXiv preprint arXiv:2310.03965*.
[14] Yu, P., Xu, J., Weston, J., & Kulikov, I. (2023). Distilling system 2 into system 1. Meta FAIR Research.
[15] Zhou, T., Chen, B., Qin, C., & Zhao, W. (2023). Evaluating large language models on multi-document summarization: Current challenges and future directions. *arXiv preprint arXiv:2310.14735*.