From Instruction to Output: The Role of Prompting in Modern NLG

Anonymous ACL submission

Abstract

Prompt engineering has emerged as an integral technique for extending the strengths and abilities of Large Language Models (LLMs) to gain significant performance gains in various Natural Language Processing (NLP) tasks. This approach, which requires instructions to be composed in natural language to bring out the knowledge from LLMs in a structured way, has driven breakthroughs in various NLP tasks. Yet, there is still no structured framework or coherent understanding of the varied prompt-engineering methods and techniques, particularly in the field of Natural Language Generation (NLG).

This brief survey aims to help fill that gap by outlining recent developments in prompt engineering, and their effect on different NLG tasks. We also position prompt design as an input-level control mechanism for NLG outputs, contrasting it with fine-tuning.

1 Introduction

The field of artificial intelligence (AI) has advanced significantly with the introduction of LLMs over the last few years. These LLMs have attained unprecedented performance in several downstream Natural Language Processing (NLP) tasks (Vatsal and Dubey, 2024), including, but not limited to, question answering, story telling, summarization, machine translation and sentiment analysis. Within this broader NLP landscape, Natural Language Generation (NLG) poses its own challenges. Despite the fluency and versatility of LLMs, generating high-quality text for diverse NLG tasks often demands careful guidance that goes beyond improvements in model architecture (Gatt and Krahmer, 2018; Pandey and Roy, 2023a) or task-specific fine-tuning (Xu et al., 2023). In this context, prompt engineering (Liu et al., 2023) has emerged as a promising paradigm for steering LLM outputs in NLG, enabling flexible control over style, structure and content, without additional retraining. Prompt engineering is a technique whereby natural language instructions, or prompts, are used to guide an LLM's behavior responses, with the aim of improving accuracy, relevance and coherence in the generated output (Chen et al., 2025). This approach offers a practical, low-resource alternative for advancing NLG applications, making it increasingly relevant as the field moves towards building robust and adaptable generation systems.

Prompt design plays a critical role in shaping the structure and coherence of a response across a wide range of NLG tasks (Schulhoff et al., 2024). Owing to its use of instructions and examples written in natural language, it constitutes a bridge between users and LLMs, letting users decide and guide an LLM's behavior. Despite its increasing popularity and significance, prompt engineering remains underrepresented in existing surveys on NLG. Most of the surveys focus on innovation in architecture designs (Pandey and Roy, 2023b; Zarrieß et al., 2021), evaluation methods (Stent et al., 2005; Sai et al., 2022; Celikyilmaz et al., 2020), or the categorization of downstream tasks (Dong et al., 2022; Gatt and Krahmer, 2018; Santhanam and Shaikh, 2019). As a result, there is a need for a survey focusing just on prompting for NLG and its applications.

In this brief survey paper, we compare prompt engineering with fine-tuning and decoding-level strategies, present a taxonomy of prompting techniques for NLG, analyze how prompts enable control over content, structure, and style, without retraining, demonstrating their utility in real-world NLG tasks, and consider emerging trends and challenges in prompt-based NLG. Our goal is to formalize prompt engineering as an emerging input-level control strategy and provide a foundation for future research in prompt-based NLG.

2 Comparison with Fine-Tuning and Decoding-Level Control

Prompt-based control operates at the input level, requiring no extra training and providing fast adaptability to new control goals. In contrast, fine-tuning allows deeper integration of control signals by aligning an LLM's internal representations with desired outputs. This achieves higher consistency than prompt-based NLG outputs, but at higher computational and data costs (Shin et al., 2025).

Decoding-level control (e.g., constrained beam search, nucleus sampling adjustments) (Naseh et al., 2023) manipulates the generation process after an LLM's next-token probabilities are computed, enabling some lexical or length constraints, but with limited flexibility for high-level aspects of generation, such as discourse structure, tone, or content framing (Holtzman et al., 2020).

Thus, prompt engineering holds a unique middle position. It is more cost-effective and adaptable than fine-tuning, while offering broader control dimensions than decoding-level control. This makes prompt engineering an effective strategy across various NLG tasks such as summarization, story generation, dialogue generation etc. (Vatsal and Dubey, 2024).

3 Taxonomy of Prompting Techniques

In this section, we briefly discuss different prompting paradigms, and how suited they are for different NLG tasks (Table 1). The core paradigms are categorized into zero-shot, few-shot, chain-of-thought, thread of thought, chain of event and role prompting.

Zero-shot Prompting (Radford et al., 2019). This technique relies on carefully-curated prompts to guide the LLMs in performing specific NLG tasks, such as machine translation (MT) and story telling. Its strength lies in the quick deployment of an idea with a relatively low design effort.

Few-shot Prompting (Brown et al., 2020). This technique leverages the idea of in-context learning to provide a few input-output pairs to improve an LLM's understanding of a given task. Eliciting even a few high-quality examples has yielded performance gains on different NLG tasks, such as dialogue generation and machine translation, steering the output towards certain stylistic nuances.

Chain-of-thought (CoT) Prompting (Wei et al., 2022). This technique takes its inspiration from how people decompose a complex task into smaller sub-tasks before arriving at the final solution. Along the same lines, this prompting paradigm provides instructions to an LLM in such a way that encourages a step-by-step, coherent reasoning process. CoT can help LLMs plan outlines, sub-points and narrative arcs for NLG tasks such as storytelling, report generation and summarization, thereby improving global coherence.

Thread-of-thought (ThoT) Prompting (Zhou et al., 2023). This technique, which draws its inspiration from human cognitive processes,

is designed to enhance the reasoning abilities of LLM models by asking them to select pertinent information from context comprising information from diverse sources, including user queries, conversation history, and external knowledge bases. ThoT has yielded substantial performance gains on tasks like conversational question answering and dialogue system where the generation of contextually appropriate answer is critical.

Chain-of-event (CoE) Prompting (Bao et al., 2024). This technique, which was proposed for summarization, consists of four steps: (1) extract specific events, (2) analyze and generalize the extracted events in more refined and concise form, (3) filter the generalized events to retain only those that cover most of the text, and (4) integrate the events selected in step (3) based on their chronological order or level of importance.

Role Prompting (Kong et al., 2023). This technique involves assigning a role to the LLM to enhance its understanding of the task. For example, if the model is prompted to act as a mathematician, it is likely to provide a correct step-by-step explanation of a mathematical concept (Van Buren, 2023). It also serves as an effective implicit CoT trigger, explaining its enhancements in reasoning capabilities.

4 Prompting for Controlled NLG

Despite the LLMs ability to generate fluent and grammatically sound texts, other control dimensions are important in real-world applications, e.g., style, content, length and structure, which might not be achievable if we allow LLMs to generate text freely (Ajwani et al., 2024). Controlling model outputs to fit a set of constraints is the purview of controllable text generation (Lu et al., 2023), and prompting serves as a *'control lever'*

Table 1: Comparative Overview of Prompting Paradigms in NLG.

Prompting Paradigm	**Primary Strategy**	**Strengths**	**Limitations**	**Typical NLG Tasks**
Zero-shot (Radford et al., 2019)	No examples; relies on pretrained generalization to generate predictions	Minimal setup cost, quick deployment	Sensitive to changes in prompt phrasing	Classification, QA, and machine translation
Few-shot (Brown et al., 2020)	In-context examples to demonstrate task pattern	Improves task-specific performance with few samples	Token budget grows with example count; brittle to formatting	Structured generation, summarization, and dialogue generation
Chain-of-Thought (Wei et al., 2022)	Step-by-step reasoning in prompt	Enhances reasoning transparency and accuracy	Verbose outputs; performance gain is task-dependent	Explainable generation, multi-turn QA, and arithmetic/logical explanation
Thread-of-Thought (Zhou et al., 2023)	Maintains context across turns using guided flow	Improves discourse and context tracking; suitable for long dialogues	Harder to automate threading; needs clear structure	Conversational QA, data-to-text generation
Chain-of-Event (Bao et al., 2024)	Extracts and compresses event chains in stages	Improves summary coherence and fluency	Narrow scope; task-specific design	Multi-document summarization
Role Prompting (Kong et al., 2023)	Assigns persona or role to guide behavior	Boosts creativity, diversity, and task framing	Relies on accurate persona crafting	Dialogue agents, story generation, creative writing

for steering the outputs of LLMs, without requiring extensive parameter fine-tuning.

4.1 Content Control: Topical Constraints and Lexical Anchoring

The limits of generative LLMs are unclear, yet from research perspective, we must be able to determine what makes them successful and what causes them to fail (Lu et al., 2023). Carefully crafted prompts can enforce topical constraints by explicitly specifying the domain or keywords to be included in the output. For example, “ Generate a short explanation about the challenges blind and low vision individuals face in accessing data visualizations in educational settings” can anchor the generation of text around the relevant topic and subtopics. Similarly, lexical anchoring, which refers to specifying important information to the LLM before reaching a decision, can guide the model to generate domain-specific information without re-training (Tian and Zhang, 2024). For example, prompting with “Write a paragraph about the challenges in web accessibility for blind users. Make sure to include the terms *tactile graphics*, *screen reader*, and *access barrier*" anchors the model in accessibility terminology and context. These approaches are useful for NLG tasks such as question answering (Zhu et al., 2021), story telling (Fan et al., 2018), and summarization (Xu et al., 2024) etc.

4.2 Structure Control: Length and Discourse Organization

Users often expect generated texts to fall within a specific length range, making length controlled generation an important topic (Jie et al., 2024). For pre-trained language models, the most widely applied technique for length control is prompt-based fine-tuning (e.g., ‘summarize in two sentences’ or ‘provide a bullet-point list of advantages of living in coastal areas) (Liu et al., 2023). In addition, prompts can also guide an LLM to simulate the discourse structure of human-written text (Ghazvininejad et al., 2022).

4.3 Style Control: Formality, Emotion, and Tone

Prompting can effectively condition the formality level (‘write in a formal academic style’ vs. ‘explain in a casual, friendly tone’), emotional coloring (‘write a comforting message to someone anxious about exams’), and persona-based tone (’as a supervisor, advise on thesis writing’). Drawing its inspiration from the impact of language on human performance, EmotionPrompt (Li et al., 2023) utilizes emotional cues to help improve an LLM’s emotional appeal. Compared to style transfer methods requiring paired data or explicit attribute modeling, prompting offers a flexible and low-resource alternative (Yang and Carpuat, 2025). In addition to written output, the emotional prompts may be used to control and guide the emotional tone of synthesized speech (Bott et al., 2024).

5 Evaluation and Prompt Robustness

5.1 Evaluation Metrics

To systematically understand the impact of prompt engineering, we need to look at different metrics that how prompting impacts the generated output’s quality and it is necessary to look at both, human-centered and automatic, metrics to capture output quality comprehensively. Both methods have their own pros and cons. Human-centered evaluation is more in-line with human intuition, but is more time-consuming and expensive (Paul et al., 2024). On the other hand, automatic evaluation is much quicker and less expensive than human-centered. However the best way to use an evaluation an LLM-generated output's quality depends on the specific application (Stent et al., 2005).

Human Evaluation: Human evaluation depend on human evaluators to assess the quality of the generated output and are often employed to capture dimensions difficult to quantify automatically such as coherence, fluency, relevance, and factuality (Holtzman et al., 2020). Human evaluations are increasingly used to assess tasks where the generated content is more abstract, such as writing and summarization (Stiennon et al., 2020; Yao et al., 2023; Wang et al., 2024).

Automatic Evaluation: Automatic evaluation uses algorithms to assess the quality of the output by an LLM, measuring the efficacy of different prompting strategies. Metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) remain widely used for surface-level text similarity, while BERTScore (Zhang et al., 2019) aims to assess at a higher semantic level. In the context of safe and responsible text generation, toxicity and bias score (Gehman et al., 2020) are significantly employed to so that the prompt formulations do not unintentionally elicit harmful content. However, these automated metrics often fail to capture the assessment results of human evaluators fully and therefore must be used with caution (Sai et al., 2022).

5.2 Prompt Sensitivity and Brittleness

LLM-generated outputs can be highly sensitive and variable to the prompt phrasing and structure, with performance often differing markedly across models and tasks based on these nuances (Zhuo et al., 2024; Chatterjee et al., 2024; Salinas and Morstatter, 2024). This sensitivity manifests in two formats: i) *local sensitivity* refers to the small lexical changes such as synonyms or reordering of instructions may lead to noticeable difference in tone, quality and relevance of the generated response. ii). *global brittleness* refers to changes in prompt length or specificity can lead to failures in adhering to the intended task, indicating a lack of robustness in prompt-based control compared to more structured fine-tuning methods.

To mitigate prompt sensitivity and brittleness, prompt tuning (Lester et al., 2021) was introduced where soft prompt vectors are learned while keeping the LLM parameters frozen. This works in a different way than manual prompt design, which allows for creativity and iterative experimentation. While manual design allows for domain-specific control with low resources, prompt tuning offers improved consistency across tasks by optimizing prompts in a differentiable manner.

6 Emerging Trends and Open Challenges

The field of prompt engineering is evolving through various key trends. *Prompting with retrieval augmented generation* (RAG) (Lewis et al., 2020) enforce factual grounding and reduce hallucinations by pairing text generation with retrieval. *Prompt engineering as programming* (Sharma et al., 2024) is capturing attentions via prompt DSL, a domain-specific language for refining prompts and monitoring the inputs to the LLM-based chatbots.

A persistent challenge is prompt's generalization and transferability across different domains, as many prompts remain fragile and task-specific (Perez et al., 2021). Also, with plenty of pre-trained LLMs to choose from, how to choose them to better leverage prompt-based learning is another interesting and difficult problem (Liu et al., 2023). Furthermore, there are fairly wide variety of tuning strategies available for prompts, LLMs, or both. However, given that this research field is at emergent stage, we still lack a systematic understanding of the tradeoffs between these methods.

7 Conclusion

Prompting has evolved into a core technique for LLM-driven NLG, enabling controlled, efficient generation without retraining. However, prompt design often involves repetitive and time-consuming debugging, as small phrasing changes can lead to unpredictable outputs.

To advance the field, prompt engineering should be formalized with frameworks, benchmarks and theory that support robust, scalable, and reusable prompt design for the next generation of NLG systems. There is also a pressing need for evaluation metrics that more accurately reflect the effectiveness and control capabilities of prompts.

8 Limitations

Despite offering a focused overview of prompt engineering for NLG, this survey has several limitations. Findings discussed are largely conceptual and may not generalize across domains, languages, or LLM architectures without further empirical validation. While we highlight issues such as prompt sensitivity and evaluation gaps, practical mitigation strategies and theoretical modeling could not be explored in depth. Future work is needed to extend the taxonomy, assess generalizability across multilingual settings, and ground prompt engineering in more rigorous empirical and theoretical frameworks.

References

Rohan Deepak Ajwani, Zining Zhu, Jonathan Rose, and Frank Rudzicz. 2024. Plug and play with prompts: A prompt tuning approach for controlling text generation. *arXiv preprint arXiv:2404.05143*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Songlin Bao, Tiantian Li, and Bin Cao. 2024. Chain-of-event prompting for multi-document summarization by large language models. *International Journal of Web Information Systems*, (ahead-of-print).

Thomas Bott, Florian Lux, and Ngoc Thang Vu. 2024. Controlling emotion in text-to-speech with natural language prompts. *arXiv preprint arXiv:2406.06406*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Survey*, 55(8).

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Marjan Ghazvininejad, Vladimir Karpukhin, Vera Gor, and Asli Celikyilmaz. 2022. Discourse-aware soft prompting for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4570–4589.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Prompt-based length controlled generation with multiple control types. *arXiv preprint arXiv:2406.10278*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1982–2008.

Ali Naseh, Kalpesh Krishna, Mohit Iyyer, and Amir Houmansadr. 2023. Stealing the decoding algorithms of language models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1835–1849.

Abhishek Kumar Pandey and Sanjiban Sekhar Roy. 2023a. Natural language generation using sequential models: A survey. *Neural Process. Lett.*, 55(6):7709–7742.

Abhishek Kumar Pandey and Sanjiban Sekhar Roy. 2023b. Natural language generation using sequential models: A survey. *Neural Processing Letters*, 55(6):7709–7742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Survey*, 55(2).

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651.

Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, and 1 others. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Reshabh K Sharma, Vinayak Gupta, and Dan Grossman. 2024. SPML: A dsl for defending language models against prompt attacks. *arXiv preprint arXiv:2402.11755*.

Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. 2025. Prompt engineering or fine-tuning: An empirical assessment of llms for code. In *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*, pages 490–502. IEEE.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. page 341–351.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Yuan Tian and Tianyi Zhang. 2024. Selective prompt anchoring for code generation. *CoRR*, abs/2408.09121.

David Van Buren. 2023. Guided scenarios with simulated expert personae: a remarkable strategy to perform cognitive work. *arXiv preprint arXiv:2306.03104*.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different NLP tasks. *CoRR*, abs/2407.12994.

Rui Wang, Hongru Wang, Fei Mi, Boyang Xue, Yi Chen, Kam-Fai Wong, and Ruifeng Xu. 2024. Enhancing large language models against inductive instructions with dual-critique prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5345–5363.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lei Xu, Mohammed Asad Karim, Saket Dingliwal, and Aparna Elangovan. 2024. Salient information prompting to steer content in prompt-based abstractive summarization. *arXiv preprint arXiv:2410.02741*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Xinchen Yang and Marine Carpuat. 2025. Steering large language models with register analysis for arbitrary style transfer. *arXiv preprint arXiv:2505.00679*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: a survey. *Information*, 12(9):355.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2023. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.