

# Reverse Prompt Engineering in Large Language Models

## A Meta-Study of Applications, Advantages, and Comparisons with Forward Techniques

Symbio6<sup>1</sup> Published: January 2025

### Abstract

The integration of Large Language Models (LLMs) into diverse fields has significantly impacted natural language processing, with prompt engineering emerging as a critical technique for optimising model outputs. This paper presents a meta-study comparing two primary approaches in this domain: Reverse Prompt Engineering (RPE) and Forward Prompt Engineering (FPE). While FPE focuses on designing structured prompts to guide LLMs toward specific outcomes, RPE analyses outputs to infer and refine the original prompts, providing flexibility in less structured situations. By systematically reviewing 15 studies, we highlight the key trends, applications, and limitations specific to RPE.

Our findings highlight RPE's effectiveness in adversarial and privacy-sensitive environments, its compatibility with black-box models, and its adaptability to dynamic tasks. Conversely, FPE excels in structured contexts requiring high efficiency and scalability. Hybrid approaches integrating RPE and FPE emerge as a promising direction, leveraging their complementary strengths to address complex, multi-turn, and security-sensitive tasks. However, significant challenges remain, including the lack of standardised evaluation metrics, computational complexity, and semantic alignment.

This study proposes a framework for assessing RPE methods, integrating accuracy, efficiency, and real-world applicability. It also underscores the ethical considerations of prompt misuse and highlights the need for further research to enhance scalability and performance in diverse applications. By providing a comprehensive evaluation of RPE and FPE, this paper advances the understanding of prompt engineering and outlines future directions for research and practice.

**Keywords:** Large Language Models (LLMs), Prompt Engineering, Reverse Prompt Engineering (RPE), Forward Prompt Engineering (FPE)

---

<sup>1</sup> Correspondence: [symbio6.nl/en/contact](https://symbio6.nl/en/contact)

## 1. Introduction

The rapid adoption of Large Language Models (LLMs) across sectors such as healthcare, law, and finance has marked a significant milestone in the evolution of Natural Language Processing (NLP). The branch of AI focused on enabling machines to understand, interpret, and generate human language. LLMs are advanced AI systems built on neural network architectures and trained on extensive datasets to execute diverse NLP tasks, including text generation, summarisation, and comprehension. By 2023, LLMs had become indispensable tools, with their utility further enhanced by advances in prompt engineering—a critical method for guiding these models to produce context-specific and relevant outputs. Prompt engineering is broadly categorised into two methods: Forward Prompt Engineering (FPE) and the emerging Reverse Prompt Engineering (RPE) [1], [2].

FPE represents a structured approach to crafting prompts that explicitly direct LLMs to generate precise and reliable outputs [3], [4]. The development of models like GPT-3 demonstrated the efficacy of FPE, which uses well-designed inputs to achieve task-specific outcomes. This approach has been pivotal in applications requiring high accuracy and consistency. However, the dynamic and evolving demands of real-world scenarios have necessitated the introduction of RPE, an innovative method that refines prompts by analysing the outputs they generate. RPE has gained recognition for its adaptability and utility in environments where understanding model reasoning is crucial, particularly in flexible AI systems that evolve with input changes and complex decision-making tasks [1], [2], [5].

RPE focuses on analysing model outputs to infer the original prompts, making it especially effective for black-box models where internal parameters are inaccessible [1], [3]. This method addresses challenges in interpreting and optimising model behaviour, offering unique advantages in scenarios requiring explainability and adaptability. While FPE has dominated practical applications due to its straightforward implementation and established methods, RPE offers a complementary approach that is still underexplored and holds significant potential for advancing prompt engineering in domain-specific contexts [5], [6], [7].

This paper presents a comparative analysis of RPE and FPE, focusing on their approaches, applications, and implications across various domains. Despite the growing interest in RPE, its performance and applicability remain underexamined compared to the well-established practices of FPE.

## 2. Methodology

A systematic review and combining data to analyse multiple studies (meta-analysis) were conducted in accordance with the standardised Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a rigorous and unbiased synthesis of existing research [S1], [S2]. Methodological refinements were systematically documented through protocol amendments, adhering to best practices for systematic reviews and addressing the specific

objectives of this study. This comprehensive approach balanced methodological rigour with adaptability, enabling a robust evaluation of emerging research on text-based prompt engineering.

The review adhered to carefully defined criteria to ensure relevance and maintain methodological transparency:

- **Types of Studies:** The analysis included peer-reviewed journal articles, conference proceedings, and grey literature. Each source was assessed using the AACODS checklist, which evaluates Authority, Accuracy, Coverage, Objectivity, Date, and Significance. Additionally, the GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) framework was applied to assess methodological rigour, ensuring consistency in comparison and inclusion.
- **Domains:** The focus was restricted to text-based prompt engineering applications, explicitly excluding vision-language models. This domain-specific scope aligned with the study's objectives, maintaining methodological relevance and focus.
- **Inclusion Criteria:** Studies were included if they utilised Reverse Prompt Engineering (RPE) or Forward Prompt Engineering (FPE) methodologies and reported measurable outcomes, such as accuracy, semantic alignment, efficiency, or security.
- **Exclusion Criteria:** Studies lacking methodological transparency or those focused on non-text-based tasks were excluded to safeguard the integrity of the review.

A comprehensive search strategy was developed to ensure the inclusion of relevant literature. Searches were conducted in Google Scholar and supplemented with Google Search (top 20 results). A snowball technique further expanded the search, tracing forward and backward citations from the initially included studies.

The search query was designed to encompass studies related to RPE and FPE while explicitly excluding vision-language models. The following search string was used:

*("reverse prompt engineering" OR "prompt inversion" OR "prompt recovery" OR "prompt deconstruction")  
AND ("large language models" OR "LLMs")  
AND ("forward prompt engineering" OR "prompt optimisation" OR "prompt design")  
NOT ("image generation" OR "vision-language models")*

ChatGPT-4o was employed to refine the search, synthesise themes, and identify literature gaps, leveraging its advanced capabilities to process large volumes of text and highlight emerging trends.

Given the heterogeneity of the included studies, data synthesis emphasised quantitative meta-analysis. To address potential gaps and broaden the analysis, a benchmarking of these results with grey literature was conducted.

The systematic review framework was selected for its structured and replicable methodology. The exclusion of vision-language models was a deliberate decision to reduce complexity and maintain

consistency across selected studies. While this focus enhances reliability, it highlights the need for future research to address multimodal applications (e.g., text, images, and audio).

Ethical considerations were integral to the research process, despite the study's reliance on secondary data and the absence of ethics board approval requirements. Transparency in data handling, respect for intellectual property, and the disclosure of potential conflicts of interest were prioritised.

While the methodology was robust, certain limitations must be acknowledged. The exclusion of vision-language models restricts the generalisability of findings to other modalities. Additionally, reliance on grey literature introduces variability in quality. This limitation was mitigated by rigorously applying the AACODS checklist to ensure the inclusion of only credible sources.

This study adheres to PRISMA guidelines and integrates supplementary quality frameworks, such as AMSTAR (A Measurement Tool to Assess Systematic Reviews, appendix A) and GRADE. Detailed documentation of protocols and amendments further enhances transparency, reproducibility, and alignment with best practices in systematic reviews [S1], [S2].

### 3. Results

#### 3.1 Selection process

##### Overview of Included Studies

The systematic review identified 53 documents, of which 15 met the inclusion criteria (Table 1). These studies represent a focused effort to evaluate Reverse Prompt Engineering (RPE) methods in domain-specific applications of Large Language Models (LLMs). The majority of the included studies were published in 2024, reflecting the nascent and rapidly evolving nature of this research area. Among these, five were peer-reviewed articles, seven were preprints, and three were grey literature sources. To ensure methodological rigour, grey literature was analysed separately. None of the included studies that combine results from multiple studies were meta-analyses or systematic reviews.

Table 1. Overview of Study Selection Process

Method	Identified Documents	Duplicates Removed	Screened	Excluded	Included
1. Google Scholar	7	1	6	4	2
2. Google Search Top20	20	4	16	10	6
3. Snowball Backwards	15	3	12	7	5
4. Snowball Forwards	11	1	10	8	2
<b>Total</b>	<b>53</b>	<b>9</b>	<b>44</b>	<b>29</b>	<b>15</b>

The exclusion reasons, summarised in Table 2, involved eliminating studies based on criteria such as irrelevance to the research question, insufficient methodological validation, or technical limitations.

Table 2. Exclusion Reasons Summary for Excluded Papers

Criteria	Number of Excluded Studies
Not relevant to RPE/FPE	14
Focused only on FPE	7
Insufficient rigour/validation	5
Technical limitations	2
Other	1
<b>Total</b>	<b>29</b>

The effectiveness of paper collection methods was assessed using three key criteria: number included, precision, and normalised Rank-Biased Precision (nRBP). Table 3 presents these metrics:

- **Precision:** Represents the proportion of included studies relative to the total identified. Snowball Backwards demonstrated the highest precision, with 42% of retrieved studies included, while Snowball Forwards had the lowest precision at 20%.
- **nRBP:** Assesses how effectively relevant studies were prioritised in the results list. Google Search achieved the best ranking performance with an nRBP score of 0.48. The lowest-ranked paper included in the Google Search Top20 results was positioned at #13.

Table 3. Metrics of Paper Collection Methods

Method	Included (#)	Accuracy (Precision)	Ranking (nRBP)
1. Google Scholar	2	0.33	0.38
2. Google Search Top20	6	0.38	0.48
3. Snowball Backwards	5	0.42	-
4. Snowball Forwards	2	0.20	-
<b>Overall</b>	<b>15</b>	<b>0.34</b>	-

A baseline query using the term “reverse prompt engineering” in Google Scholar yielded four results, two of which were included in the study, achieving a precision of 50%. While this is relatively high, it identified fewer studies compared to the 15 included papers obtained through the hybrid search strategy.

### Domain Coverage and Bias Analysis

The included studies spanned diverse domains, such as healthcare, legal systems, creative writing, and industrial applications, highlighting the flexibility of RPE for use in different industries.

However, generalisability to other contexts or domains was limited by the dependence on task-specific datasets and proprietary data.

A risk of bias assessment revealed that five studies demonstrated high methodological rigour with robust datasets and transparent reporting. Conversely, three studies exhibited low rigour due to vague methodologies and reliance on restricted-access data. Publication bias, characterised by the disproportionate publication of research with significant results, was particularly evident in smaller studies, highlighting the importance of cautious interpretation and the need for replication.

### **Variability in Metrics and Ethical Considerations**

Significant methodological variability complicated cross-study comparisons. Evaluation metrics such as BLEU and cosine similarity were inconsistently applied. Furthermore, the studies utilised a range of LLMs, with ChatGPT from OpenAI being the most frequently used (7 studies), followed by Meta's Llama (6 studies) and Mistral (3 studies).

Ethical considerations also varied: some studies addressed risks like prompt misuse thoroughly [5], [3], while others overlooked critical ethical concerns [2]. This heterogeneity poses challenges to synthesising findings and limits their applicability across contexts.

## **3.2 Used Implementations**

### **Overview of Implementation Methods**

Among the 12 studies reviewed, 11 detailed their implementation methods, all of which were compatible with black-box models. This compatibility eliminates the need for access to internal model parameters, distinguishing these approaches from older white-box techniques, such as Logit2Prompt, which rely on insights into internal workings like logits [1]. The black-box compatibility underscores the flexibility of RPE methods, especially in contexts where model internals are inaccessible.

### **Categorisation of RPE Implementation Scenarios**

The studies employed RPE in three primary scenarios, showcasing its adaptability across a range of tasks, including security analyses, performance optimisation, and adaptive systems (Table 4):

1. **Output-Only RPE:** Relies exclusively on outputs to deduce prompts. This method is commonly applied in adversarial tasks, which involve testing or exploiting AI model vulnerabilities and in reconstructing restricted-access prompts.
2. **RPE with Input-Output Pairs:** Utilises both inputs and outputs for iterative refinement, a process of step-by-step prompt improvement through repeated adjustments to achieve better alignment with desired outputs and enhanced optimisation.
3. **RPE Using Partial Information:** Combines known inputs with observed outputs, making it suitable for adaptive or privacy-sensitive contexts.

Table 4. Reverse Prompt Engineering (RPE) Implementation Types and Applications

Implementation Type	Key Features	Applications	References
Output-Only	Deduces prompts solely from outputs	Security, adversarial analysis, explainability	[1], [2], [6], [8]
With Input-Output Pairs	Analyses inputs and outputs iteratively	Performance optimisation, compliance	[3], [4], [9], [10]
Partial Information	Leverages known inputs and observed outputs	Adaptive systems, traceability	[5], [7], [11], [12]

### Generalised Step-by-Step Method for RPE

A generalised method for RPE implementation emphasises iterative refinement and optimisation, aligning prompts with desired outputs. The steps include:

1. **Analyse Output:** Examine tone, structure, style, and context to infer likely prompts [1].
2. **Generate Candidate Prompts:** Use heuristics, reverse engineering, or embedding-based methods to formulate candidate prompts [1], [5].
3. **Iterative Refinement:** Refine prompts by adjusting keywords, structure, and tone [6]; employing genetic algorithms and feedback loops [1]; or reducing uncertainty with probabilistic techniques [7].
4. **Test Prompts:** Validate candidate prompts by comparing outputs for similarity and accuracy.
5. **Optimise Prompts:** Remove unnecessary elements and reduce ambiguity to streamline prompts [3], [2].
6. **Select Final Prompt:** Choose the best-performing prompt based on accuracy and output similarity.

### Additional Implementation Enhancements

Several studies introduced novel techniques that enrich the RPE process:

- **Embedding-Based Inference:** Improves prompt recovery accuracy by comparing output embeddings with candidate prompts in high-dimensional vector spaces [5], [4].
- **Sparse Encoder Setup:** Efficiently extracts key patterns and features from input-output mappings using sparse encoder architectures—streamlined models designed to detect patterns with minimal computation—thereby optimising the process with compact designs [5].
- **Contextual Hints:** Leverages contextual clues such as keywords, sentence structure, or tone for iterative refinement [7].
- **Uncertainty Quantification:** Utilises probabilistic methods to identify high-uncertainty areas in generated prompts, enabling targeted refinement for greater precision [7].
- **Data Augmentation:** Enhances model generalisation by training with semi-synthetic datasets, improving performance across diverse tasks [2].

### 3.3 Metrics and Evaluation

The evaluation of RPE methods utilised a variety of metrics to assess performance, robustness, and practicality (Table 5). Accuracy and semantic alignment emerged as the most commonly applied criteria. For instance, [1] demonstrated the efficiency of RPE by recovering high-quality prompts using as few as five outputs. Beyond these metrics, robustness was also a priority, with studies evaluating models' resilience to adversarial inputs and noisy environments.

Table 5. Key Criteria and Metrics for Reverse Prompt Engineering (RPE) Evaluation

Criterion	Definition	Metric	Studies
Accuracy	Precision in recovering original prompts	BLEU, cosine similarity	[1], [2], [5], [11]
Alignment	Semantic alignment between prompts/outputs	Semantic coherence	[6], [11]
Efficiency	Outputs required for recovery	Number of outputs	[1]
Robustness	Resilience to adversarial/noisy inputs	Token sensitivity	[3], [10]
Fluency	Closeness to natural language distribution	Perplexity	[2], [3]
Security	Risk of exposing sensitive prompts	Prompt leakage	[4]

### 3.4 Applications and Use Cases

The studies reviewed highlight the versatility of RPE across a range of domain-specific applications, demonstrating its adaptability to diverse technical and operational challenges:

- **Security:** RPE safeguards sensitive data in high-stakes domains like legal and medical systems by reconstructing prompts from outputs, reducing risks of data leakage. [5] highlights its robust protection in sensitive environments, while [3] emphasises its role in isolating sensitive information to maintain security.
- **Explainability:** Enhancing the clarity of how AI systems generate their results is a key application of RPE, improving the interpretability of language model outputs. Studies such as [1] and [6] demonstrate how RPE reconstructs prompts to reveal the latent instructions that guide LLMs, thereby elucidating the reasoning behind their outputs.
- **Performance Optimisation:** RPE has been effectively applied to enhance large-scale AI deployments by iteratively reconstructing and refining prompts. As demonstrated in [1], RPE aligns model outputs with specific operational objectives, improving task efficiency and overall performance. Similarly, [2] highlights RPE's role in leveraging iterative optimisation strategies to achieve higher accuracy and efficiency in text transformation tasks.

### 3.5 RPE vs. FPE Comparison

The reviewed studies provide a detailed comparison of RPE and Forward Prompt Engineering (FPE), highlighting their respective strengths and limitations:

- **Advantages of RPE:** RPE exhibits exceptional flexibility in unstructured environments, such as multi-turn conversations between users and the AI system or adversarial scenarios, where predefined prompts are less effective [5], [3]. Its robustness against adversarial attacks and its ability to operate with limited contextual information make it particularly valuable in high-security and dynamic tasks.
- **Advantages of FPE:** FPE excels in structured, predefined tasks that require consistent and efficient outputs. With fewer computational demands than RPE, FPE is better suited for scalable applications where repeatability and predictability are critical [1], [2]. Its ability to leverage well-optimised prompts ensures high performance in static contexts.
- **Limitations:** While RPE demonstrates adaptability, its performance diminishes in noisy or ambiguous data environments, where prompt recovery becomes less precise [5], [3]. Conversely, FPE struggles to adapt to dynamic or multi-turn tasks, limiting its applicability in unstructured scenarios [1], [5]. These limitations underscore the complementary nature of RPE and FPE, suggesting that hybrid approaches could combine their strengths to overcome individual shortcomings.

### 3.6 Hybrid Models and Synergies

Emerging hybrid approaches demonstrate the potential for integrating the strengths of RPE and FPE to address complex challenges. For example, [1] discusses iterative workflows that combine RPE's adaptability for refining prompts in dynamic environments with FPE's consistency in structured tasks, enabling models to handle both predefined and evolving objectives. Though not extensively explored in the reviewed studies, such approaches offer promising solutions for bridging the gap between structured and unstructured tasks.

### 3.7 Grey Literature Insights

The core principles and outcomes of RPE discussed in the academic papers are mirrored in the practical examples provided within the grey literature [A], [B], [C]. This examination of grey literature, specifically from [A], reveals a notably broader application of RPE. This source showcases RPE's use in diverse fields such as marketing, branding, and technical tasks like code snippet generation, which are less emphasised in academic papers. This highlights RPE's practical adaptability and potential for application beyond the conventional academic research domains.

## 4. Discussion

### 4.1 Key Trends and Gaps

This meta-study highlights the limited availability of high-quality research on Reverse Prompt Engineering (RPE), revealing significant gaps in the literature. While 2024 witnessed a surge in publications in this emerging field, the absence of prior meta-studies or systematic reviews underscores the need to consolidate findings and critically assess evolving trends. The review identified substantial variability in methodologies, evaluation metrics, and domain applications,

complicating the synthesis of findings and limiting generalisability. To address these challenges, this study emphasised qualitative data synthesis, which provides a broader perspective on RPE's development and applications.

## 4.2 Search Strategies

The analysis of search strategies reveals important trade-offs between precision, relevance, and breadth. For instance, the Snowball Backwards method achieved the highest precision (0.42), Google Search effectively prioritising relevant studies (0.48). These findings underscore the importance of hybrid search strategies that combine precision-orientated methods with relevance-orientated approaches to achieve comprehensive and high-quality literature reviews. The analysis also indicated that grey literature offered limited added value in this context. Future reviews should adopt hybrid strategies to leverage the strengths of different methods while addressing their individual weaknesses.

## 4.3 Model Representation Bias

The representation of Large Language Models (LLMs) in the reviewed studies diverged significantly from real-world usage. For example, Mistral was disproportionately represented in three studies, whereas widely used models from Google and Anthropic were under-represented, appearing in only one study each. This disparity may reflect publication bias, where researchers favour more accessible models. To enhance the relevance and applicability of findings, future research should prioritise the inclusion of commercially available models that are widely used in practical applications, thereby addressing this representational imbalance.

## 4.4 Metric Standardisation

The evaluation of RPE methodologies suffers from inconsistent metrics, which limits cross-study comparability. Among 12 reviewed papers, 11 employed multiple metrics but often omitted critical measures of effectiveness. Supplementary sources the systematic review suggest that traditional metrics like BLEU and cosine similarity, which focus on surface-level matching, are becoming less useful [S3], [S4]. Instead, measures like semantic coherence and multidimensional evaluation are seen as better at capturing deeper alignment between prompts, beyond just text matching [S3], [S4]. None of the studies included in the review mention this observation.

Despite this progress, significant gaps remain in evaluating the practical utility of RPE. Effectiveness metrics such as human evaluation, task success rate, and efficiency-related metrics (e.g., outputs needed, iterations required, resource costs, learnability) are vital for assessing real-world applicability. To advance the field, future research should prioritise metrics that capture both semantic and functional alignment while integrating measures of robustness and usability. Standardising evaluation practices will enhance cross-study comparability and improve the overall impact of RPE methodologies.

## 4.5 Implementation and Applications

The results highlighted three primary implementation approaches for RPE: output-only methods, input-output pairs, and partial information approaches. These methods demonstrate RPE's

adaptability across a range of applications, including security, explainability, and performance optimisation.

Table 6 illustrates the fundamental differences between RPE and Forward Prompt Engineering (FPE) across six key steps. While their methodologies diverge significantly in the early stages, they converge in areas such as refinement, optimisation, and final prompt selection.

RPE’s flexibility allows it to excel in unstructured and adversarial environments, while FPE thrives in predefined, structured tasks. These contextual differences emphasise that the choice between RPE and FPE depends on task complexity, resource availability, and security requirements. RPE’s compatibility with black-box models makes it ideal for confidential or sensitive applications, while FPE’s efficiency is suited for scalability in structured contexts.

#### 4.6 Hybrid Approaches

Hybrid approaches hold significant promise for improving both performance and scalability in applications where standalone RPE or FPE methodologies may fall short. These models combine RPE’s flexibility with FPE’s structured efficiency, enabling their application across a broader range of tasks. Future work should focus on assessing the scalability and computational efficiency of these hybrid methods to maximise their practical applicability in both structured and unstructured scenarios.

Table 6. Differences Between Reverse Prompt Engineering (RPE) and Forward Prompt Engineering (FPE) by Step

Step	Difference	RPE vs. FPE
1. Analyse Output	Major	RPE analyses existing outputs to infer prompts; FPE begins with defining goals.
2. Generate Prompts	Major	RPE relies on reverse engineering; FPE uses creative, task-driven methods.
3. Iterative Refinement	Moderate	RPE focuses on reconstruction fidelity; FPE emphasises alignment with goals.
4. Test Prompts	Major	RPE validates prompts through output similarity; FPE tests for task-specific quality.
5. Optimise Prompts	Moderate	RPE optimises for accuracy; FPE focuses on usability and adaptability.
6. Select Final Prompt	Minimal	Both select the best-performing prompt, though evaluation criteria differ.

#### 4.7 Ethical and Practical Considerations

Ethical concerns surrounding RPE, such as the potential misuse of prompt reconstruction to expose sensitive data or confidential information, remain a significant challenge. While methods like obfuscation—hiding or masking sensitive data—and API safeguards have been explored [3], [10],

comprehensive frameworks are needed to mitigate these risks effectively. Privacy-sensitive domains like healthcare and legal systems require robust defences to ensure data integrity and confidentiality.

Moreover, insights from grey literature, such as RPE's applications in marketing and branding [A], highlight the need to align ethical considerations with practical implementations. Balancing innovation with accountability will be crucial to ensuring the responsible use of RPE methodologies in diverse contexts.

## 5. Conclusions

This study underscores the complementary strengths of Reverse Prompt Engineering (RPE) and Forward Prompt Engineering (FPE) in addressing diverse challenges in prompt design and optimisation. RPE demonstrates significant effectiveness in dynamic, unstructured, and domain-specific contexts, offering advantages in adaptability, explainability, and compatibility with black-box models. However, its reliance on high-quality data and computational intensity necessitates careful deployment, along with the implementation of robust ethical safeguards to prevent misuse. Conversely, FPE excels in structured and static workflows, where its efficiency and scalability are particularly advantageous, although it struggles to adapt to dynamic and evolving scenarios.

The integration of RPE's flexibility with FPE's structured efficiency in hybrid approaches represents a promising avenue for addressing their respective limitations. Hybrid methodologies have the potential to achieve broader applicability across dynamic and complex domains, particularly in tasks that demand a balance between adaptability and robustness. Such approaches are especially relevant in areas requiring multi-turn interactions, real-time adaptability, and security-sensitive applications.

Future research should focus on developing robust hybrid methodologies, integrate ethical safeguards against misuse, and expand these techniques to noisy, multimodal, and cross-domain datasets. Including commercially available models used in real-world applications will enhance relevance and address representational imbalances. Standardized evaluation metrics and transparent reporting are also essential for promoting comparability and trust in this emerging field.

By addressing these areas, RPE and FPE approaches can evolve to meet the growing demands of real-world applications, fostering innovation and trust in this emerging field.

## 6. Acknowledgments

The authors gratefully acknowledge the financial support of Symbi6, which has enabled the successful completion of this study.

## References

- [1] H. Li and D. Klabjan, "Reverse prompt engineering," *preprint*, arXiv:2411.06729v2, Nov. 2024. Available: [arxiv.org/abs/2411.06729v2](https://arxiv.org/abs/2411.06729v2).
- [2] J. Chen, W. Xu, Z. Ding, J. Xu, H. Yan, and X. Zhang, "Advancing prompt recovery in NLP: A deep dive into the integration of Gemma-2b-it and Phi2 models," *preprint*, arXiv:2407.05233v1, Jul. 2024. Available: [arxiv.org/abs/2407.05233v1](https://arxiv.org/abs/2407.05233v1).
- [3] Y. Yang et al., "PRSA: Prompt stealing attacks against large language models," *preprint*, arXiv:2402.19200v2, Jun. 2024. Available: [arxiv.org/abs/2402.19200v2](https://arxiv.org/abs/2402.19200v2).
- [4] Y. Zhang, N. Carlini, and D. Ippolito, "Effective prompt extraction from language models," presented at the Conf. on *Computational Linguistics and Models (COLM)*, 2024. DOI: 10.36227/techrxiv.123456789/v1.
- [5] C. Zhang, J. X. Morris, and V. Shmatikov, "Extracting Prompts by Inverting LLM Outputs," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, FL, USA, 2024, pp. 14753-14777. Association for Computational Linguistics. Available: <https://aclanthology.org/2024.emnlp-main.819/>.
- [6] Z. Sha and Y. Zhang, "Prompt stealing attacks against large language models," *preprint*, arXiv:2402.12959v1, Feb. 2024. Available: [arxiv.org/abs/2402.12959v1](https://arxiv.org/abs/2402.12959v1).
- [7] L. Gao, R. Peng, Y. Zhang, and J. Zhao, "DORY: Deliberative prompt recovery for LLM," *preprint*, arXiv:2405.20657v2, Jun. 2024. Available: [aclanthology.org/2024.findings-acl.631.pdf](https://aclanthology.org/2024.findings-acl.631.pdf).
- [8] H. Li, M. Xu, and Y. Song, "Sentence Embedding Leaks More Information Than You Expect: Generative Embedding Inversion Attack to Recover the Whole Sentence," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada, Jul. 2023, pp. 14022–14040. Available: [aclanthology.org/2023.findings-acl.881/](https://aclanthology.org/2023.findings-acl.881/)
- [9] L. Give, T. Zaoral, and M. A. Bruno, "Uncovering hidden intentions: Exploring prompt recovery for deeper insights into generated texts," *preprint*, arXiv:2406.15871v1, Jun. 2024. Available: [arxiv.org/abs/2406.15871v1](https://arxiv.org/abs/2406.15871v1).
- [10] D. Agarwal et al., "Prompt leakage effect and defense strategies for multi-turn LLM interactions," in *Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing: Industry Track*, Nov. 2024, pp. 1255–1275. Available: [github.com/salesforce/prompt-leakage](https://github.com/salesforce/prompt-leakage).
- [11] S. Femepid, L. Hatherleigh, and W. Kensington, "Gradual improvement of contextual understanding in large language models via reverse prompt engineering," *preprint*, Aug. 2024. Available: [doi.org/10.22541/au.172376001.14254079/v1](https://doi.org/10.22541/au.172376001.14254079/v1).
- [12] H. Duan, A. Dziedzic, M. Yaghini, N. Papernot, and F. Boenisch, "On the privacy risk of in-context learning," *preprint*, arXiv:2411.10512v1, Nov. 2024. Available: [arxiv.org/abs/2411.10512v1](https://arxiv.org/abs/2411.10512v1).

*Grey Literature:*

[A] Anonymous, "LLM prompt recovery: Recover the prompt used to transform a given text," [Discussion Post]. Kaggle, Apr. 2024. Available: [www.kaggle.com/competitions/llm-prompt-recovery/discussion/494755](https://www.kaggle.com/competitions/llm-prompt-recovery/discussion/494755). Accessed: Dec. 24, 2024.

[B] S. Uspenskyi, "Comprehensive guide to reverse prompt engineering: Exploring types, use cases, and examples," Springs AI Knowledge Hub, Feb. 2024. Available: [springsapps.com/knowledge/comprehensive-guide-to-reverse-prompt-engineering---all-you-need-to-know](https://springsapps.com/knowledge/comprehensive-guide-to-reverse-prompt-engineering---all-you-need-to-know). Accessed: Dec. 24, 2024.

[C] H. Mohapatra, "Reverse prompt engineering: A deep dive with examples," LinkedIn, Oct. 2024. Available: [www.linkedin.com/pulse/reverse-prompt-engineering-deep-dive-examples-dr-hitesh-mohapatra-koy7c](https://www.linkedin.com/pulse/reverse-prompt-engineering-deep-dive-examples-dr-hitesh-mohapatra-koy7c). Accessed: Dec. 24, 2024.

*Supplementary:*

[S1] "Meta-Study Protocol: Systematic Review and Meta-Analysis of Artificial Intelligence Methodologies," symbio6.nl, Dec. 23, 2024. Available: [symbio6.nl/p/meta-study-protocol-ai-methodologies.pdf](https://symbio6.nl/p/meta-study-protocol-ai-methodologies.pdf).

[S2] "Protocol Amendments for the Systematic Review on Reverse and Forward Prompt Engineering," symbio6.nl, Dec. 23, 2024. Available: [symbio6.nl/p/meta-study-protocol-amendments-rpe-llm.pdf](https://symbio6.nl/p/meta-study-protocol-amendments-rpe-llm.pdf).

[S3] Y. Liu, Y. Su, E. Shareghi, and N. Collier, "Unlocking Structure Measuring: Introducing PDD, an Automatic Metric for Positional Discourse Coherence," *preprint* preprint arXiv:2402.1017, 2024. Available: [arxiv.org/abs/2402.10175](https://arxiv.org/abs/2402.10175)

[S4] K. Iida and K. Mimura, "CATER: Leveraging LLM to Pioneer a Multidimensional, Reference-Independent Paradigm in Translation Quality Evaluation," *preprint*, arXiv:2412.11261, 2024. Available: [arxiv.org/abs/2412.11261](https://arxiv.org/abs/2412.11261)

## Appendix A: AMSTAR Assessment

This appendix presents the detailed AMSTAR (A Measurement Tool to Assess Systematic Reviews) assessment of the meta-study, evaluating the quality of the systematic review and identifying areas for improvement.

### Summary of AMSTAR Assessment

Domain	Score
Research Questions Clearly Defined	Yes
Comprehensive Literature Search	Partially
Duplicate Selection and Extraction	Yes
Search Strategy Documentation	Yes
Study Inclusion Justified	Yes
Risk of Bias Assessment	Partially
Methods for Combining Findings	Yes
Publication Bias Considered	Yes
Conflict of Interest Statement	Yes
Study Validity Assessed	Partially
Funding Sources	No
Replication Details	Yes

### Overall Rating: High Quality

This systematic review demonstrates strong adherence to AMSTAR guidelines in most domains. The primary areas for improvement include the inclusion of formal bias assessment tools, funding disclosures, and expanded database coverage. Addressing these gaps will further enhance the reliability and comprehensiveness of the findings.

### *Authors' Comment on Partially Graded Scores*

*The partially graded scores in the AMSTAR assessment stem from the inherent heterogeneity of the included studies. Given this variability, the focus of data synthesis was directed toward a quantitative meta-analysis where feasible, supplemented by qualitative synthesis to address gaps. Additional databases for the literature search yielded no extra relevant results, emphasising the focused scope and emerging nature of the research field.*