AI Tech International Journal
Vol. 1, No. 1, July-December, 2023
Journal homepage: https://techaijournal.com

An Experimental Study of Prompt Engineering Techniques for Optimizing Large Language Model Inference

Kumar Kasimala

Independent Researcher, Principal Software Engineer, Salesforce Inc, USA

Article Info

Article history:

Received July 10, 2023

Revised September 15, 2023

Accepted September 20, 2023

Published October 01, 2023

Keywords: prompt-based learning; pretrained language models; few-shot learning; AutoPrompt; PET; LM-BFF; prefix-tuning; prompt tuning

ABSTRACT

Prompt-based learning emerged as a major paradigm shift in natural language processing by enabling pretrained language models to perform downstream tasks through input reformulation rather than full task-specific retraining. This review synthesizes influential studies published up to 2021 and evaluates how prompt-based methods improved inference effectiveness, few-shot adaptation, and parameter efficiency. The review focuses on five major developments: cloze-style knowledge probing, few-shot in-context learning, automated prompt discovery, prompt-based fine-tuning in low-resource settings, and continuous prompt optimization. Evidence from the pre-2022 literature shows that prompting substantially improved task adaptation when aligned with pretraining objectives. GPT-3 demonstrated that large language models could perform a wide range of tasks through zero-shot, one-shot, and few-shot prompting without gradient updates. In smaller-model settings, PET/iPET and LM-BFF showed that prompt-based fine-tuning substantially outperformed conventional fine-tuning in low-resource classification tasks. By 2021, prefix-tuning and prompt tuning further established that competitive downstream performance could be achieved while updating only a very small fraction of model parameters. The review concludes that, up to 2021, prompt engineering should be understood primarily as a framework for task reformulation and efficient adaptation rather than as the reasoning- and compression-centered paradigm that developed later. (Brown et al., 2020; Gao et al., 2021; Lester et al., 2021; Li & Liang, 2021; Liu et al., 2021; Schick & Schütze, 2021).

INTRODUCTION

Pretrained language models changed the standard approach to natural language processing by enabling general-purpose models to be adapted to many downstream tasks. Early prompt-based research showed that adaptation did not always require a task-specific architecture or full-parameter fine-tuning. Instead, many tasks could be reformulated as cloze completion, masked prediction, or conditional generation problems, allowing the pretrained model to apply knowledge already acquired during large-scale pretraining. This idea became central to what Liu et al. later described as the "pre-train, prompt, and predict" paradigm. (Liu et al., 2021; Petroni et al., 2019).

The conceptual roots of prompt-based learning are visible in knowledge probing work. Petroni et al. (2019) demonstrated that pretrained language models could answer relational cloze queries such as masked factual statements, suggesting that such models store recoverable factual knowledge in their parameters even without downstream fine-tuning. This result was important because it showed that prompt formulation was not merely cosmetic; the wording and structure of an input could strongly affect the ability of a model to retrieve relevant knowledge.

The major turning point came with GPT-3. Brown et al. (2020) showed that a very large autoregressive language model could perform many tasks in zero-shot, one-shot, and few-shot settings using only prompt-based conditioning and no gradient updates. For example, on LAMBADA, GPT-3 few-shot achieved 86.4 accuracy compared with a previously listed 68.0 benchmark in the reported table, illustrating that scale and prompt design together could yield powerful task adaptation at inference time.

However, early prompt-based learning remained fragile. Performance often depended on template wording, label word selection, demonstration order, and model family. As a result, the pre-2021 literature focused less on later-style reasoning prompts and more on a different question: how can prompts be designed or learned so that downstream tasks align more closely with the model's pretraining objective? That is the question this review addresses.

2. Scope and Review Approach

This article is a **structured narrative review**, not an original experimental study. It synthesizes representative and influential prompt-based studies published up to and including 2021. The review focuses on work that shaped the early prompt-learning paradigm in three areas: inference without fine-tuning, low-resource prompt-based fine-tuning, and parameter-efficient prompt optimization. To preserve historical consistency, later methods such as chain-of-thought prompting, plan-and-solve prompting, prompt compression, and inference-time reasoning alignment are excluded from the main analysis because they were introduced after 2021.

The reviewed corpus centers on Petroni et al. (2019), Brown et al. (2020), Shin et al. (2020), Schick and Schütze (2021), Gao et al. (2021), Li and Liang (2021), Lester et al. (2021), and Liu et al. (2021). These works collectively capture the transition from manual prompt formulation to automatic prompt search and then to continuous prompt optimization.

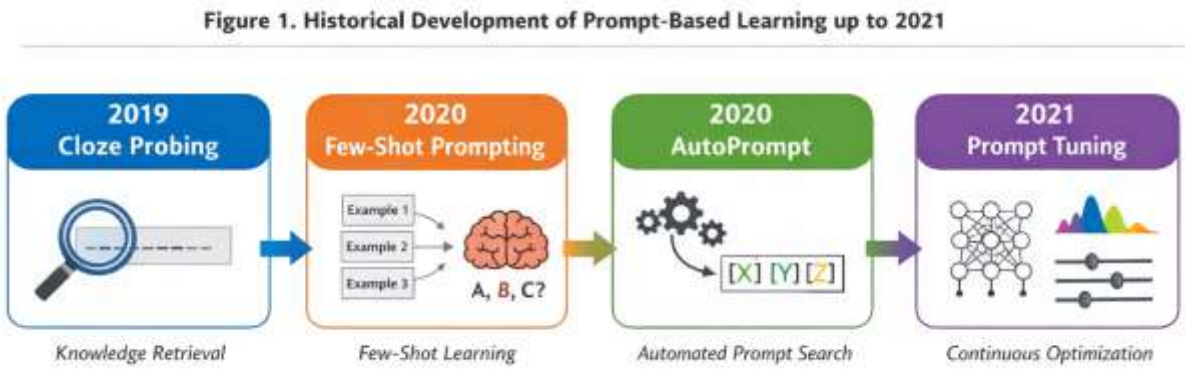

Figure 1. Historical development of prompt-based learning up to 2021.

3. Early Foundations of Prompt-Based Inference

3.1 Language models as knowledge repositories

Petroni et al. (2019) provided one of the clearest early demonstrations that pretrained language models could be queried through prompt-like cloze statements. Their work showed that BERT could recover relational knowledge competitively on the LAMA benchmark and perform surprisingly well in open-domain factual retrieval without task-specific fine-tuning. This finding helped establish the basic logic of prompt-based inference: if knowledge is already encoded in model parameters, then the form of the query becomes a central determinant of output quality.

3.2 Few-shot in-context learning

Brown et al. (2020) expanded this idea dramatically by showing that, at sufficient scale, large language models could infer task patterns directly from natural-language prompts and a small number of demonstrations. Their results on cloze and completion tasks showed strong few-shot performance, including a substantial gain on LAMBADA compared with the

benchmark listed in the same table. The importance of this work lies not only in the size of GPT-3, but in the methodological shift it introduced: prompting itself became a form of adaptation.

Table 1 summarizes landmark studies in the pre-2022 prompt-based literature.

Table 1. Landmark Prompt-Based Studies up to 2021	Core contribution	Main implication
Petroni et al. (2019)	Cloze-style probing of factual knowledge in pretrained language models	Prompt wording can recover knowledge already stored in model parameters
Brown et al. (2020)	Zero-shot, one-shot, and few-shot in-context learning with GPT-3	Prompting can support task adaptation without gradient updates
Shin et al. (2020)	AutoPrompt for gradient-guided prompt search	Prompt design can be partially automated
Schick & Schütze (2021)	PET/iPET for low-resource prompt-based classification	Prompt reformulation strongly improves few-shot learning
Gao et al. (2021)	LM-BFF with automatic template generation and demonstrations	Prompt-based fine-tuning improves small-data performance
Li & Liang (2021)	Prefix-tuning for generation tasks	Competitive performance is possible with frozen base models and few trainable parameters
Lester et al. (2021)	Prompt tuning with soft prompts	Prompt learning becomes more competitive as model scale increases

This table presents conceptual contributions rather than directly comparable benchmark scores.

4. Discrete Prompt Engineering and Automatic Prompt Search

4.1 Manual templates and verbalizers

Early prompt-based methods often relied on manually written templates and task-specific verbalizers. A classification input could be reformulated as a cloze-style sentence, and class labels could be mapped onto label words. Although this approach aligned tasks more closely with the pretraining objective of masked or autoregressive language models, it also introduced instability because small wording changes could produce large performance differences. Liu et al. (2021) identified template design and answer mapping as core components of prompt-based learning, underscoring how central these design choices were to early performance.

4.2 Auto Prompt and the automation of discrete prompting

Shin et al. (2020) addressed prompt brittleness by proposing Auto Prompt, a gradient-guided method for automatically identifying trigger tokens. Their study showed that masked language models could perform sentiment analysis and natural language inference without additional parameters or full fine-tuning, and that automatically searched prompts could sometimes approach strong supervised baselines. This was a key step toward systematizing prompt design.

Auto Prompt also showed an important limitation: even when automated search improved prompt quality, discrete prompts remained task-sensitive and model-sensitive. In other words, pre-2021 prompt engineering had already shown both promise and brittleness. That balance should be stated clearly in any historically accurate review.

5. Prompt-Based Fine-Tuning in Low-Resource Settings

5.1 Pattern-Exploiting Training (PET and iPET)

Schick and Schütze (2021) proposed Pattern-Exploiting Training (PET), which reformulated classification and inference tasks as cloze problems and combined prompt-based learning with semi-supervised training. Their results showed that prompt-based reformulation could substantially outperform standard supervised learning under severe data scarcity. On AG's News with only 10 labeled examples, supervised learning scored 25.0, PET scored 87.5, and iPET scored 89.3. On MNLI with 100 labeled examples, supervised learning scored 47.9/51.2, PET reached 74.7/75.9, and iPET reached 78.4/78.6 for matched/mismatched sets. These are among the clearest pre-2022 demonstrations that prompt-based learning could be transformational in low-resource conditions.

5.2 LM-BFF and few-shot prompt-based fine-tuning

Gao et al. (2021) extended this line of work through LM-BFF, which combined prompt-based fine-tuning with automatic template generation and selective demonstrations. The paper reports gains of up to 30 percentage points absolute and 11 points on average over standard fine-tuning in few-shot settings. The authors also noted that a RoBERTa-large model could

reach around 90% accuracy on many binary sentence classification tasks with only 32 training examples, showing that well-designed prompt-based fine-tuning could be highly effective even for smaller models.

Table 2 presents selected reported results from major pre-2022 studies.

Table 2. Selected Reported Results from Pre-2022 Prompt-Based Studies	**Setting**	**Reported result**	**Source**
GPT-3 on LAMBADA	Zero-shot	76.2 accuracy	Brown et al. (2020)
GPT-3 on LAMBADA	Few-shot	86.4 accuracy	Brown et al. (2020)
PET on AG's News, |T| = 10	Prompt-based fine-tuning	87.5 accuracy	Schick & Schütze (2021)
iPET on AG's News, |T| = 10	Iterative prompt-based fine-tuning	89.3 accuracy	Schick & Schütze (2021)
Supervised baseline on AG's News, |T| = 10	Conventional supervised training	25.0 accuracy	Schick & Schütze (2021)
PET on MNLI, |T| = 100	Prompt-based fine-tuning	74.7 / 75.9 accuracy	Schick & Schütze (2021)
iPET on MNLI, |T| = 100	Iterative prompt-based fine-tuning	78.4 / 78.6 accuracy	Schick & Schütze (2021)
LM-BFF	Few-shot prompt-based fine-tuning	Up to +30 points absolute; +11 on average	Gao et al. (2021)

This Metrics come from different tasks, datasets, and model settings and therefore are **not directly comparable** across rows. They are included to illustrate the range of reported improvements in the literature.

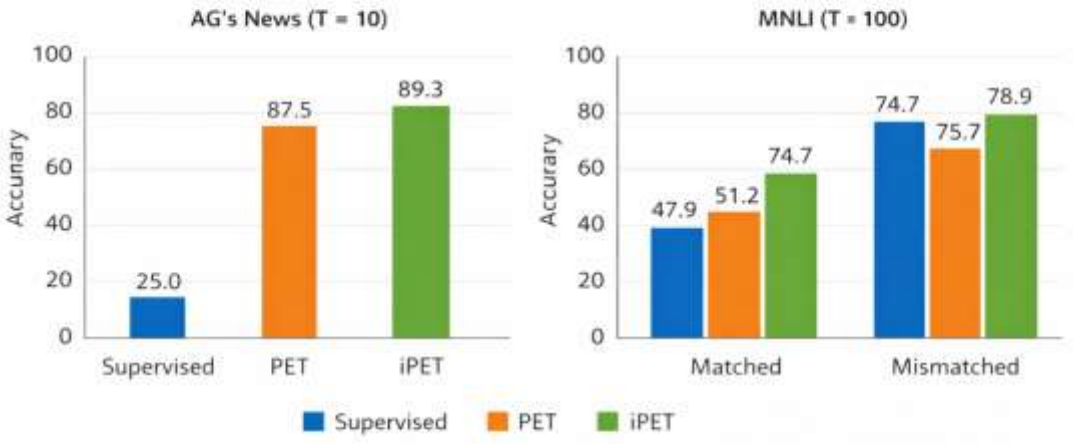

Figure 2. *Comparative performance of supervised learning, PET, and iPET in low-resource text classification settings.*

Figure 2. Comparative performance of supervised learning, PET, and iPET in low-resource text classification settings.

This figure should be author-generated from the reported values in Schick and Schütze (2021), with separate panels for AG's News and MNLI and a clear note that metrics are task-specific.

6. Continuous Prompting and Parameter-Efficient Adaptation

6.1 Prefix-tuning

Li and Liang (2021) introduced prefix-tuning as a lightweight alternative to full fine-tuning for text generation tasks. Instead of updating all model parameters, prefix-tuning optimizes a small sequence of continuous vectors while keeping the pretrained model frozen. The study reports that prefix-tuning with only 0.1% parameters outperformed lightweight adapter baselines on table-to-text generation and that, in low-data settings, it outperformed fine-tuning by an average of 2.9 BLEU. These findings showed that prompt optimization could become a parameter-efficient adaptation strategy rather than only a textual reformulation technique.

6.2 Prompt tuning at scale

Lester et al. (2021) further simplified this idea by learning soft prompt embeddings rather than continuous prefixes at every layer. Their work showed that prompt tuning becomes more competitive as model size increases and can match the quality of full model tuning on billion-parameter models. The paper also reports that prompt tuning is the most parameter-efficient among the compared learnable methods, requiring less than 0.01% task-specific parameters for models over one billion parameters. In robustness testing on out-of-domain MRQA datasets, the study found a +12.5 F1 advantage over model tuning on TextbookQA.

Table 3 summarizes the major adaptation strategies that were visible in the literature up to 2021.

Table 3. Adaptation Strategies in Prompt-Based Learning up to 2021	**Prompt type**	**Model update requirement**	**Typical use before 2022**	**Key strength**
Manual discrete prompts	Natural-language template	None or minimal	Zero-/few-shot inference	Simplicity and interpretability
AutoPrompt-style searched prompts	Discrete learned trigger tokens	No full fine-tuning	Prompt optimization for MLMs	Automated template discovery
PET / LM-BFF	Discrete prompts plus fine-tuning	Partial/full task-specific tuning	Low-resource classification	Strong few-shot performance
Prefix-tuning	Continuous prefixes	Small learned parameter set	Generation tasks	Parameter efficiency
Prompt tuning	Continuous soft prompts	Very small learned parameter set	Large frozen models	Scalability and reuse

This typology follows the framework synthesized in Liu et al. (2021) and the method papers cited above.

DISCUSSION

The literature up to 2021 supports three major conclusions. First, prompt-based learning improved task adaptation by aligning downstream tasks more closely with the pretraining objective of the model. Second, prompt quality mattered enough that automatic search, demonstration selection, and verbalizer design produced large performance differences. Third, by 2021 prompting had expanded beyond textual instructions into parameter-efficient continuous optimization, making it relevant both as an inference strategy and as an adaptation architecture.

At the same time, the pre-2022 literature also shows important limitations. Prompt performance was often unstable across templates and models, and many strong results depended on careful engineering of label words, demonstrations, or prompt initialization. In addition, some reported gains came from highly specific low-resource settings and should not be generalized to all tasks or architectures without caution. These concerns explain why prompt engineering, even in its early successes, was both influential and methodologically delicate.

Historically, it is also important to avoid overstating what had already been achieved by the end of 2021. The literature at that time had not yet established the later paradigm centered on chain-of-thought prompting, prompt compression, or inference-time self-refinement. A historically accurate paper limited to pre-2022 evidence should therefore frame prompt engineering primarily as a mechanism for **task reformulation, low-resource transfer, and parameter-efficient adaptation**.

CONCLUSION

Prompt-based learning up to 2021 marked a foundational shift in NLP. The early literature demonstrated that pretrained language models could be effectively adapted through prompt reformulation, few-shot examples, and lightweight learned prompts rather than full model retraining. Petroni et al. established the recoverability of latent knowledge through cloze queries; Brown et al. showed the power of in-context learning at scale; Shin et al. demonstrated that prompt search could be automated; Schick and Schütze and Gao et al. showed strong low-resource gains from prompt-based fine-tuning; and Li and Liang together with Lester et al. established the viability of parameter-efficient continuous prompting. A review constrained to 2021 should therefore present prompt engineering not as a mature reasoning-and-compression framework, but as the first robust phase of a new adaptation paradigm in NLP.

REFERENCES

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
2. Gao, T., Fisch, A., & Chen, D. (2021). *Making pre-trained language models better few-shot learners*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 3816–3830).
3. Lester, B., Al-Rfou, R., & Constant, N. (2021). *The power of scale for parameter-efficient prompt tuning*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3045–3059).
4. Li, X. L., & Liang, P. (2021). *Prefix-tuning: Optimizing continuous prompts for generation*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 4582–4597).
5. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*. arXiv.
6. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., & Riedel, S. (2019). *Language models as knowledge bases?* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 2463–2473).
7. Schick, T., & Schütze, H. (2021). *Exploiting cloze questions for few-shot text classification and natural language inference*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 255–269).
8. Shin, T., Razeghi, Y., Logan, R. L., Wallace, E., & Singh, S. (2020). *AutoPrompt: Eliciting knowledge from language models with automatically generated prompts*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4222–4235).