# Language Models are Few-Shot Learners - LAMBADA dataset

Abisherk Sivakumar

Department of Computer Science and Engineering, University of Moratuwa
Colombo, Sri Lanka
abisherks.21@cse.mrt.ac.lk

**Abstract.** The LAMBADA dataset presents a challenging benchmark for language models, requiring broad discourse context to predict the final word of passages. While recent work has demonstrated that larger models achieve better performance, the role of prompt engineering and example selection remains underexplored. In this work, we investigate the impact of semantic few-shot example selection and cloze-style prompt formatting on LAMBADA performance using GPT-3.5-turbo. Our experiments demonstrate substantial improvements over random example selection, achieving 87.2% accuracy in the few-shot setting with both semantic selection and cloze prompting—a 14% absolute improvement over few-shot learning without these techniques. These results suggest that careful prompt engineering can significantly enhance language model performance on challenging completion tasks, even with smaller context windows and API token limitations.

**Keywords:** Language Models · Few-Shot Learning · Prompt Engineering · LAMBADA · Semantic Similarity

## 1 Introduction

The LAMBADA dataset [1] tests language models' ability to model long-range dependencies by requiring prediction of the final word in passages that demand broader discourse understanding. The original GPT-3 paper [2] reported significant performance gains with model scaling, achieving 86.4% accuracy with the 175B parameter model in the few-shot setting using a cloze-style format.

However, two key aspects remain relatively unexplored: (1) the impact of semantically relevant example selection versus random selection, and (2) the combined effects of cloze-style prompting and semantic selection with more accessible models. In this work, we address these gaps using GPT-3.5-turbo, a more accessible alternative to the discontinued GPT-3 API.

Our contributions include:

- Systematic evaluation of semantic example selection using sentence embeddings
- Ablation studies isolating the effects of cloze prompting and semantic selection

– Demonstration that prompt engineering techniques can partially compensate for model size differences

## 2 Methodology

### 2.1 Dataset and Model

We use the LAMBADA test set [1] and evaluate 500 examples due to API token rate limitations. For few-shot learning, we use the validation set as our source of examples. We employ GPT-3.5-turbo-instruct as our base model, which provides completion-style API access similar to the original GPT-3 models.

### 2.2 Semantic Example Selection

Following recent work in meta-learning [3], we hypothesize that semantically similar examples improve in-context learning. We implement semantic selection using the `mxbai-embed-large-v1` sentence transformer model [4] to compute embeddings for all validation examples. For each test instance, we select the $k$ most similar examples based on cosine similarity of their embeddings.

### 2.3 Cloze-Style Prompting

We adopt the cloze-style format from Brown et al. [2], which explicitly indicates the completion task:

```
[context] _____ → [target_word]
```

This format signals to the model that exactly one word should follow, avoiding the continuation bias observed in standard language modeling formats.

### 2.4 Experimental Setup

We evaluate six conditions combining presence/absence of cloze prompting and semantic selection in both one-shot ($k = 1$) and few-shot ($k = 15$) settings. All experiments use temperature=0 and top_p=1.0 for deterministic generation. The model is instructed to output exactly one word, with post-processing to remove punctuation and convert to lowercase for matching.

## 3 Results

Table 1 presents our main results across all experimental conditions.

Our results demonstrate clear improvements from both interventions. In the one-shot setting, cloze prompting alone improves accuracy by 6.2 percentage points (61.0% → 67.2%), while adding semantic selection provides an additional

**Table 1.** Accuracy on 500 LAMBADA test examples under different conditions. Both cloze-style prompting and semantic example selection contribute to improved performance, with the combination achieving the best results.

| Setting | Cloze | Semantic | Accuracy (%) |
|---|---|---|---|
| One-shot | × | × | 61.0 |
| One-shot | | × | 67.2 |
| One-shot | | | **69.2** |
| Few-shot ($k = 15$) | × | × | 73.2 |
| Few-shot ($k = 15$) | | × | 82.4 |
| Few-shot ($k = 15$) | | | **87.2** |

2.0 points (67.2% → 69.2%). The gains are more substantial in the few-shot setting, where cloze prompting improves accuracy by 9.2 points (73.2% → 82.4%), and semantic selection adds another 4.8 points (82.4% → 87.2%).

Notably, our best result (87.2%) closely approaches the GPT-3 175B performance (86.4%) reported by Brown et al., despite using a significantly smaller model. This suggests that strategic prompt engineering can partially compensate for differences in model capacity.

## 4 Analysis and Discussion

### 4.1 Effect of Semantic Selection

The consistent improvements from semantic selection (2.0% in one-shot, 4.8% in few-shot) support the hypothesis that contextually relevant examples enhance in-context learning. The larger effect in few-shot settings suggests that semantic coherence becomes more important as more examples are provided, allowing the model to better identify the task pattern. This aligns with findings in meta-learning literature [5] that task-relevant examples facilitate faster adaptation.

### 4.2 Effect of Cloze Prompting

Cloze-style prompting shows substantial benefits (6.2% in one-shot, 9.2% in few-shot), confirming observations from Brown et al. [2]. The format effectively constrains the model to produce single-word completions rather than continuations, addressing a fundamental mismatch between the task format and standard language modeling objectives. The larger improvement in few-shot settings suggests the model better learns the task format when multiple cloze-formatted examples are provided.

### 4.3 Synergistic Effects

The combination of both techniques yields superadditive improvements in the few-shot setting (14.0% total improvement vs. 9.2% + 4.8% individual effects

when measured separately). This suggests that semantic coherence and task formatting interact positively, with semantically similar examples helping the model better understand the cloze task format.

### 4.4 Limitations

Our evaluation is limited to 500 examples due to API rate constraints, which may affect statistical significance. Additionally, we use GPT-3.5-turbo rather than the original GPT-3 models, potentially introducing confounding factors. The validation set used for example selection may have different characteristics than the test set, though LAMBADA's design minimizes this concern. Future work should validate these findings on the full test set and across multiple model families.

## 5 Related Work

The related works are as follows:

**Few-shot learning:** Brown et al. [2] demonstrated that language models can perform tasks from few examples without gradient updates. Our work extends this by systematically studying example selection strategies.

**Prompt engineering:** Recent work has shown that prompt formatting significantly impacts model performance [2]. We contribute empirical evidence for the combined effects of formatting and example selection.

**Meta-learning:** The use of semantic similarity for example selection draws inspiration from matching networks [5] and other meta-learning approaches that leverage task similarity.

## 6 Conclusion

We demonstrate that semantic few-shot example selection and cloze-style prompting significantly improve language model performance on the LAMBADA dataset. Our best configuration achieves 87.2% accuracy on 500 test examples, approaching the performance of much larger models through careful prompt engineering. These results highlight the importance of example selection strategies and task formatting in few-shot learning, suggesting that substantial performance gains are possible even with fixed model architectures.

Future work should explore these techniques across other instances of the benchmarks and investigate the theoretical foundations of why semantic similarity improves in-context learning. Additionally, studying the interaction between model scale and prompt engineering could reveal whether these techniques provide consistent benefits across model sizes or primarily benefit smaller models.

## Code Availability

The code for reproducing these experiments is available at: `https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210018B-NLP_Language-Understanding`

## References

1. Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. *The LAMBADA dataset: Word prediction requiring a broad discourse context.* arXiv preprint arXiv:1606.06031, 2016.
2. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. *Language models are few-shot learners.* Advances in Neural Information Processing Systems, 33:1877–1901, 2020.
3. Chelsea Finn, Pieter Abbeel, and Sergey Levine. *Model-agnostic meta-learning for fast adaptation of deep networks.* In International Conference on Machine Learning, pages 1126–1135. PMLR, 2017.
4. Sentence Transformers library. *mixedbread-ai/mxbai-embed-large-v1.* Available at: `https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1`, 2024.
5. Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. *Matching networks for one shot learning.* Advances in Neural Information Processing Systems, 29, 2016.