Progress Evaluation

Improving One-Shot Performance and Zero-Shot Task Constraint Understanding in LAMBADA

Index Number : 20018B

1. Introduction

Language models like GPT-3 have shown remarkable performance improvements on the LAMBADA dataset through few-shot prompting, especially when the task is framed as a cloze test. However, challenges remain in the one-shot and zero-shot settings. In one-shot, cloze-style formatting sometimes performs worse than zero-shot, indicating difficulty in generalizing from a single example. In zero-shot, models often fail to infer implicit task constraints (like predicting only the final word of a passage). This project aims to address these two gaps by developing better prompting strategies for one-shot performance and enabling models to understand task constraints in zero-shot settings.

2. Objectives

- **Objective A:** Improve one-shot performance on LAMBADA by designing and testing advanced prompting strategies.

- **Objective B:** Develop methods that allow language models to correctly infer and follow task constraints in zero-shot scenarios without explicit demonstrations.

3. Research Gaps

1. **Improving One-Shot Performance with Specific Formatting:** Existing cloze-style formatting works in few-shot but fails in one-shot. A gap exists in developing prompting methods that help models adapt to this pattern from only one demonstration.

2. **Zero-Shot Understanding of Task Constraints:** Models lack awareness that only the last word of the passage is the expected completion. Few-shot examples help, but in zero-shot settings, task constraints remain poorly understood.

4. Methodology

Phase 1: Literature Review & Problem Refinement (Aug – Sep 2025)

- Review prior studies on LAMBADA performance, cloze-test framing, and prompting.
- Study advances in instruction-tuning, meta-prompting, and constraint-aware decoding.
- Refine research questions based on identified weaknesses in one-shot and zero-shot performance.
- The LAMBADA dataset has been widely used to benchmark language models on long-range dependency reasoning, with GPT-3 achieving strong improvements in the few-shot setting but still underperforming in one-shot contexts where cloze-style formatting even reduced performance compared to zero-shot (Brown et al., 2020). Earlier work suggested that scaling models and data might yield diminishing returns for LAMBADA (Kaplan et al., 2020), yet GPT-3's 18% gain in few-shot accuracy challenged this notion, highlighting the need for deeper study of scaling laws versus architectural innovations. One promising direction is parameter-efficient fine-tuning with adapter modules (Houlsby et al., 2019), which insert small bottleneck layers into transformer blocks while freezing the base model. Adapters have been shown to match full fine-tuning performance on benchmarks like GLUE while training less than 3% of parameters, and they support modular, task-specific adaptation without catastrophic forgetting. Unlike prompt-tuning or prefix-tuning, which operate at the input level, adapters modify internal representations, making them particularly suited to capturing structural constraints such as LAMBADA's requirement of predicting only the final word of a passage. Recent work on instruction-tuning (Wei et al., 2022) has also demonstrated that explicit task descriptions improve zero- and one-shot generalization, suggesting that adapters combined with cloze-style or instruction-based data could improve robustness to formatting in one-shot LAMBADA, while also enforcing task-specific constraints in zero-shot settings. Together, these studies motivate the exploration of adapter-based strategies as a lightweight but powerful alternative to brute-force scaling for improving performance on LAMBADA.

Phase 2: Dataset Preparation & Baseline Establishment (Sep 2025)

- Load LAMBADA dataset using Hugging Face datasets library.

- Establish baseline results:

    - Zero-shot performance.
    - One-shot cloze-style performance.
    - Few-shot (reference) performance.

- Conduct error analysis to understand common failure modes.

Phase 3: Strategy Development (Sep – Oct 2025)

- **For One-Shot:**
    - Implement instruction-based prompting (e.g., “Fill in the blank with exactly one word.”).
    - Test hybrid prompting (one demonstration + meta-instruction).
    - Explore prompt ensembling (multiple one-shot formats aggregated).
    - Introduce contrastive demonstrations (show one correct and one incorrect example).
- **For Zero-Shot:**
    - Apply constraint-aware decoding (restrict completions to one token or until punctuation).
    - Add task-aware natural language instructions in prompts.
    - Train with synthetic cloze datasets to induce implicit constraint awareness.
    - Explore meta-prompting (“This is a cloze test. Predict exactly one word to complete the sentence.”).

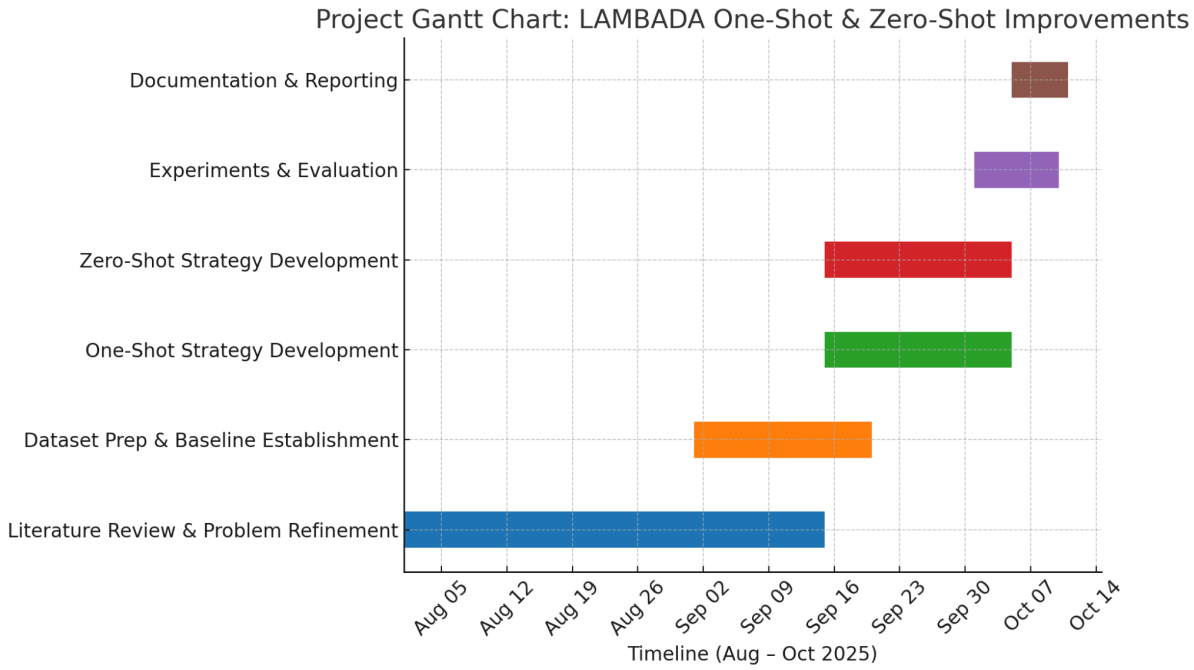
Phase 4: Experiments & Evaluation (Oct 2025)

- Run experiments on LAMBADA test set.
- Metrics: Accuracy, qualitative error analysis.
- Compare one-shot and zero-shot improvements against baselines.
- Identify best-performing strategies.

Phase 5: Documentation & Reporting (Oct 2025)

- Compile experimental results.
- Draft research findings and analysis.
- Prepare final report with methodology, results, and conclusions.
- Suggest future research directions (generalization to other cloze datasets).

5. Timeline (Gantt Chart)



6. Deliverables

- Baseline results for zero-shot, one-shot, and few-shot LAMBADA.
- Improved prompting methods for one-shot performance.
- Zero-shot constraint inference strategies.
- Experimental evaluation and analysis.
- Final project report and timeline documentation.

7. Expected Impact

This project will provide new insights into how large language models handle one-shot and zero-shot learning under strict task constraints. By addressing weaknesses in formatting sensitivity and implicit constraint recognition, the findings could generalize to other benchmarks involving structured completions, enhancing task-agnostic LM performance.

8. References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners*. In Advances in Neural Information Processing Systems (NeurIPS).
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). *Parameter-efficient transfer learning for NLP*. In Proceedings of the 36th International Conference on Machine Learning (ICML).

- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv preprint arXiv:2001.08361.
- Lester, B., Al-Rfou, R., & Constant, N. (2021). *The power of scale: Parameter-efficient adaptation for pre-trained language models*. arXiv preprint arXiv:2104.08691.
- 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 (ACL).
- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2022). *Finetuned language models are zero-shot learners*. arXiv preprint arXiv:2109.01652.