

# Supplementary Material

## Prospective Training and Post-Training Interventions for Faster OQA

The main paper shows that larger language models typically require fewer turns to solve Optimal Question Asking (OQA) tasks. Scaling is therefore an obvious lever, yet it is not the only one. When model size must remain fixed, several strategies can still reduce dialog length while preserving overall capability.

First, we can increase effective capacity without raising per-token compute by adopting sparsely gated transformers or other parameter-efficient mixture-of-experts designs such as the Switch Transformer [Fedus et al., 2021]. Reinforcement learning on simulated OQA episodes is another option. By rewarding information gain and penalizing redundant questions, the model can learn policies that finish dialogs sooner, although such focused training must be balanced against possible losses in broad generalization. In practice, proximal policy optimization [Schulman et al., 2017] or related algorithms provide a stable gradient signal, and human preference data can further shape the policy [Ouyang et al., 2022].

A second avenue involves bilevel or meta-gradient fine-tuning that places an outer objective on dialog length while keeping an inner loop of standard likelihood learning. Early results suggest that coupling gradient signal to query efficiency improves planning skills with only modest computational cost. Offline search can also be traded for online speed. Budget-constrained Tree-of-Thought or Monte Carlo search run at training time, followed by distillation of those trajectories, lets the deployed model carry a “plan in advance” capability within a single forward pass [Yao et al., 2023].

Curriculum schedules add complementary benefits. Starting with small, unambiguous synthetic tables and gradually introducing larger or more ambiguous candidate pools helps the model internalize efficient partitioning heuristics before confronting full-scale benchmarks, echoing the principles of curriculum learning [Bengio et al., 2009]. Practical tool or memory augmentation can also help. A lightweight scratchpad or retrieval module that tracks eliminated candidates mitigates the tendency of transformer decoders to forget earlier constraints in long dialogs.

None of these ideas are mutually exclusive; combining moderate scaling, search-to-distill training, and preference-aligned fine-tuning may approach oracle-level efficiency while preserving the model’s general versatility.

## References

- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion-parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601*, 2023.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009.