**CacheDistill: Efficient Query Routing for Large Language Models via Semantic Ground Truth Caching and Edge Distillation**

**Abstract**

Large language models (LLMs) are powerful but expensive to run at scale due to high memory and compute requirements. This paper proposes a practical hybrid architecture that improves inference efficiency without sacrificing output quality.

Our method combines two key ideas:

1. **Semantic Ground Truth Caching**: Queries are semantically embedded and matched against a cache of previous inputs. If a semantically similar query exists, the system retrieves the associated ground truth and generates a fresh answer from it, rather than the full model rerunning end-to-end inference.
2. **Topic-Aware Edge Distillation**: A smaller edge-hosted distilled model handles most queries. This model can be dynamically updated to specialize in domains frequently queried by a given user. For example, if a user frequently asks about *Stardew Valley*, the edge system distills relevant sub-models from the cloud LLM and preloads them locally.

This architecture reduces redundant computation, increases responsiveness, and provides a natural pathway for fine-tuned user experience personalization.