perplexity

Give conetne in pdf

Here is the content from the PDF "Decoding Speculative Decoding":

- **Title**: Decoding Speculative Decoding
- **Authors**: Minghao Yan, Saurabh Agarwal, Shivaram Venkataraman (University of Wisconsin-Madison)
- **Abstract**:
  Speculative decoding is a widely used technique to speed up inference for large language models (LLMs) without sacrificing output quality. A smaller draft model generates speculative tokens, then the target LLM verifies those tokens. The speedup depends heavily on the draft model. This paper presents over 350 experiments with LLaMA-65B and OPT-66B, showing that speculative decoding performance depends more on draft model latency than on its actual language modeling ability. The authors design new hardware-efficient draft models for speculative decoding, achieving over 111% higher throughput than existing models, generalizing across LLaMA model series and fine-tuned models.

Key Points from the Paper:

- Speculative decoding combines a draft model (small, fast) and a target model (large, accurate).
- The draft model generates several tokens, which the target validates. Matching tokens are accepted; mismatches are recalculated by the target.
- Main bottleneck: **Draft model latency** (not quality). Depth is a major cause of latency.
- Redesigning draft models (wider and shallower) boosts throughput, as verified in scaling studies using LLaMA and OPT series.
- Empirically, draft models with high accuracy on language tasks do not necessarily perform better as draft models in speculative decoding.

Results:

- Draft model latency and architecture dominate speedup, not accuracy.
- Pruned, wider draft models achieved >100% throughput improvements over standard models.
- Pruned models also reduced KV-cache size and allowed larger batch size for inference.
- The new approach generalizes to LLaMA-3 and other supervised/fine-tuned models.

Conclusion:
Designing draft models for minimal latency—not just for accuracy—can double LLM inference

throughput with speculative decoding. The findings help guide future LLM serving and system design choices.

If you want the **full text** or a more detailed section (such as benchmarks, ablation studies, or specific tables/figures), let me know which part you'd like extracted.