Verifiable & Private Inference

Methods beyond ZK & FHE

Erhan Tezcan

Lead Developer @ Dria

4.9.2025

Introduction

whoami

- Lead Developer at **Dria**
- We are building a peer-to-peer network of

Artificial Intelligence (AI), and Large Language Models (LLMs) in particular, are revolutionizing the world. Close to %10 of the entire world population is using ChatGPT alone[1].

New models are coming out every week, smashing the existing records on numerous benchmarks, with an ever increasing performance demand.[2].

We are actually progressing faster than we though we were, as noted by powerhouse's such as OpenAI, Anthropic, and Google DeepMind[3].

---

[1]https://backlinko.com/chatgpt-stats

[2]https://hai.stanford.edu/ai-index/2025-ai-index-report

[3]https://80000hours.org/agi/guide/when-will-agi-arrive/

Within this talk, we are specifically interested in the **inference** part of the AI/LLM stack. Inference is the process of running a trained model to make predictions or generate outputs based on new input data, provided in a simplified figure Figure 1.
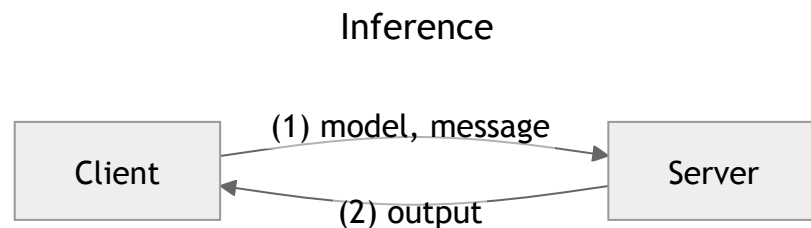
Inference

```
                    (1) model, message
  ┌────────┐  ─────────────────────────▶  ┌────────┐
  │ Client │                              │ Server │
  └────────┘  ◀─────────────────────────  └────────┘
                       (2) output
```

Figure 1: Inference

There are two problems with this figure:

-

Bibliography

Thank You!