SOVEREIGN: To what extent does the integration of secure multi-party computation protocols affect the zero-shot text clas

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Transformer models (e.g., Bert and GPT) have shown their dominance in machine learning tasks. Many cloud companies have begun to provide services based on Transformer models, examples include translation and text-speech conversion. However, such a service inevitably requires access to the client's data, which might contain sensitive information. Theoretically, running the services under secure multi-party computation (MPC) could protect clients' privacy. However, current MPC frameworks are still limited in terms of model performance, efficiency, deployment, and functionality, especially when f

1 Introduction

Analysis of: PUMA: Secure Inference of LLaMA-7B in Five Minutes. Research goal: To what extent does the integration of secure multi-party computation protocols affect the zero-shot text classification accuracy on the SetFit/llm-benchmark-suite?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

2 papers retrieved. 18 claims extracted, 1 verified. Tribunal: 4.0/10 $\rightarrow$ REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
PUMA achieves secure inference of LLaMA-7B in five minutes.	✓	0.20
PUMA was evaluated on Bert-Base, Roberta-Base, Bert-Large, GPT2-Base, GPT2-Medium, GPT2-Large, and LLaMA-7B.	×	0.05
Bert model performance was measured on the CoLA, RTE, and QNLI datasets from the GLUE benchmarks.	×	0.06
GPT2 model performance was measured on the Wikitext-103 V1 dataset.	×	0.05
The evaluation environment used Ubuntu 20.04.6 LTS with Linux kernel 5.4.0-144-generic.	×	0.01
The CPU used for evaluation operated at 2.70GHz.	×	0.01
The network bandwidth in the evaluation environment was approximately 5Gbps.	×	0.01
The round trip time in the evaluation environment was approximately 1ms.	×	0.01
PUMA's Bert-Base performance on CoLA was 0.613, compared to 0.616 for the CPU baseline.	×	0.02
PUMA's Bert-Base performance on RTE was 0.700, matching the CPU baseline of 0.700.	×	0.02
PUMA's Bert-Base performance on QNLI was 0.916, matching the CPU baseline of 0.916.	×	0.02
PUMA's Bert-Large performance on CoLA was 0.690, compared to 0.686 for the CPU baseline.	×	0.02
PUMA's Bert-Large performance on RTE was 0.747, compared to 0.755 for the CPU baseline.	×	0.02
PUMA's Bert-Large performance on QNLI was 0.918, compared to 0.922 for the CPU baseline.	×	0.02
MPCFORMER does not support loading pre-trained transformer models.	×	0.10
MPCFORMER does not implement LayerNorm faithfully.	×	0.02
MPCFORMER with Quad approximations requires retraining the modified models.	×	0.05
PUMA does not require retraining.	×	0.01

References

- https://arxiv.org/abs/2307.12533
- https://www.semanticscholar.org/paper/e382d9bddd6b160d918d386e9778df4a9d804e0c