Privacy-Preserving Machine Learning as a Service: Challenges and Architectures in Cloud Environments

Prof. Dr. Thomas Schuster

Technische Hochschule Ulm

Berufungsvortrag – THU, 2025

Agenda

Motivation

Architectural Patterns

Utility-Privacy Trade-Offs
- Needle-in-the-Haystack Testing
- Evaluation Pipeline: Utility-Privacy Benchmarking

Conclusion & Research Agenda

Motivation

- Growing adoption of ML services across industries.
- **Challenges**
  - Sensitive **data exposure**, GDPR & AI Act compliance
  - High **resource consumption** $\rightarrow$ sustainability concerns
- Need for verifiable privacy-preserving, scalable architectures.

Key Architectural Patterns

- Controlled execution environments
  (private, community, and public cloud, on-premises)

Key Architectural Patterns

- Controlled execution environments
  (private, community, and public cloud, on-premises)
- Hybrid AI Services:
  Local sensitive data processing & cloud compute outsourcing.

Key Architectural Patterns

- Controlled execution environments
  (private, community, and public cloud, on-premises)
- Hybrid AI Services:
  Local sensitive data processing & cloud compute outsourcing.
- Benchmarking utility vs. privacy trade-offs.

Hybrid AI Services

- Reduce data exposure risks & optimize resource usage.
- Process sensitive data locally.
- Use cloud for training or compute-heavy inference.

Hybrid AI Services

- Reduce data exposure risks & optimize resource usage.
- Process sensitive data locally.
- Use cloud for training or compute-heavy inference.

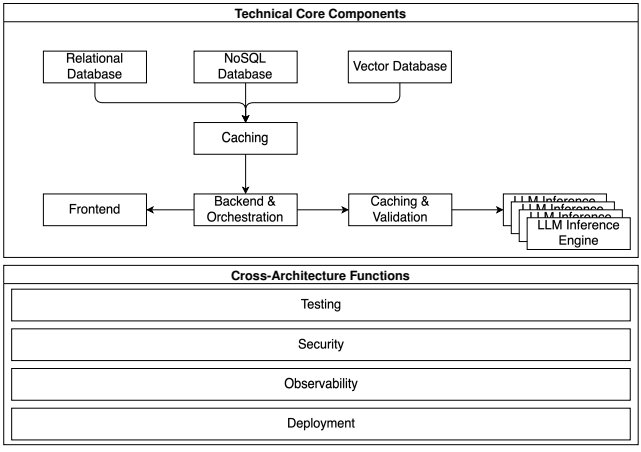
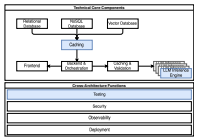


Figure: Privacy-compliant LLM deployment - Reference architecture [1]

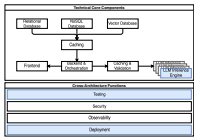
Individualization in Closed LLM Systems

- **Competitive Advantage**
  Tailored services may improve market position.
- **Efficiency Gains**
  - Integration of domain-specific terminology and custom formats.
  - Reduced need for manual post-processing [2].
- **Seamless IT Integration**
  - Custom data sources and specialized algorithms embedded into the system.
  - Differentiation in data-driven markets.



Computational Challenges and Scalability

- **Resource Requirements**
  - High demand for specialized hardware (GPUs, TPUs).
  - Local optimization needed in closed systems.
- **Optimization Strategies** [1]
  - Quantization, model compression, and Small Language Models (SLMs).
  - Efficient fine-tuning (e.g., Parameter-Efficient Fine-Tuning, PEFT).
- **Scalability**
  - Limited elastic scaling compared to public cloud systems.
  - Caching and parallel processing to ensure low latency and stable performance.



Needle-in-the-Haystack Testing

- **Problem**
  Existing benchmarks test retrieval, not reasoning.
- **Goal**
  Evaluate complex reasoning over long contexts [4].

.

Needle-in-the-Haystack Testing

- **Problem**
  Existing benchmarks test retrieval, not reasoning.
- **Goal**
  Evaluate complex reasoning over long contexts [4].
- **RQ1**
  How does context length and needle position affect performance?

.

Needle-in-the-Haystack Testing

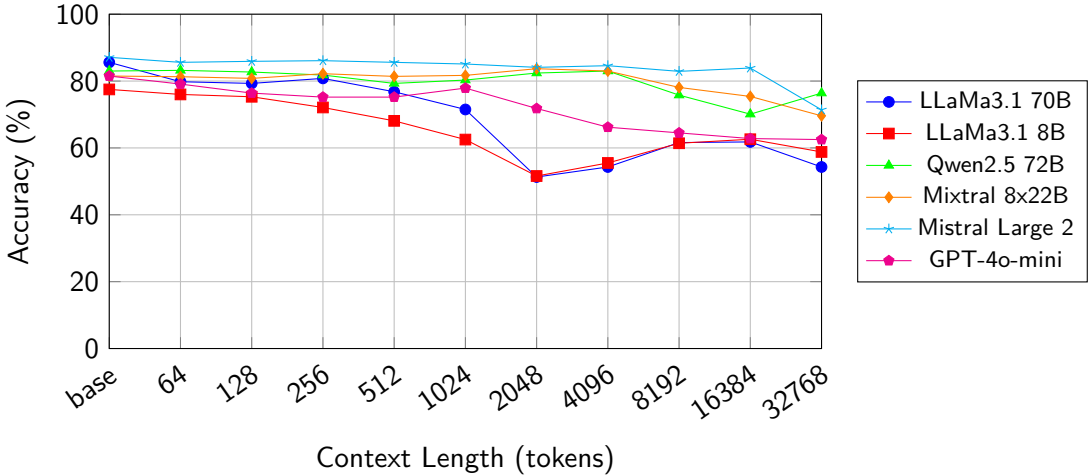
- **Problem**
  Existing benchmarks test retrieval, not reasoning.
- **Goal**
  Evaluate complex reasoning over long contexts [4].
- **RQ1**
  How does context length and needle position affect performance?
- **RQ2**
  How do different models (LLMs) compare on this task?

.

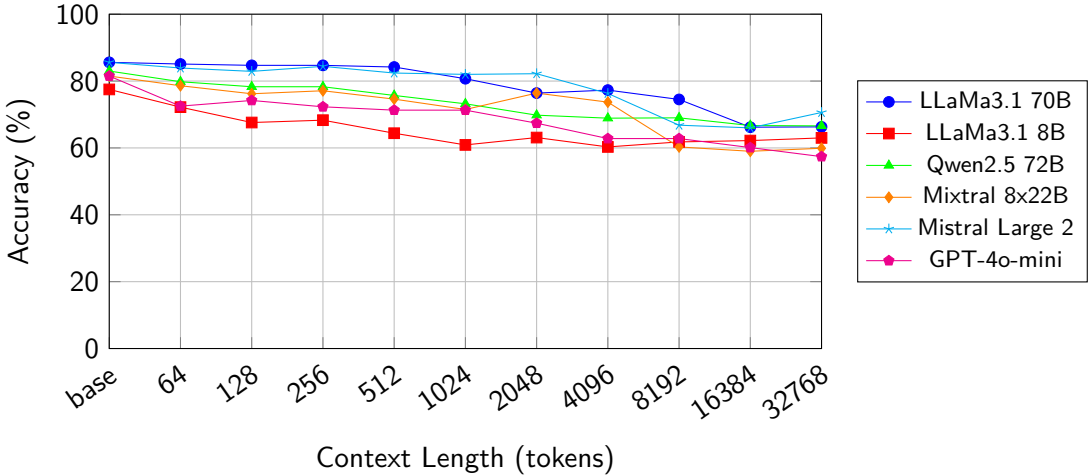
Experimental Setup

- **Task**: Verify if product descriptions violate cease-and-desist declarations.
- **Data**: 411 examples, synthetic variation, context lengths up to 32k tokens
- **Models**: LLaMa 3.1 70B & 8B, Nemotron 70B, Qwen2.5, Mixtral, Mistral Large 2, GPT-4o mini
- **Evaluation Metrics**
  - **Accuracy**: Overall proportion of correct predictions.
  - Precision (macro), Recall (macro), F1-score (macro)
  - Precision (weighted), Recall (weighted):, F1-score (weighted)

Accuracy vs. Context Size (Needle at Beginning)



Accuracy vs. Context Size (Needle at Middle)



Key Results and Conclusion

- **Findings**
  Performance drops significantly beyond 2k tokens; early/middle needles suffer most.
- **Best Model**
  Mistral Large 2 maintained robustness across context lengths.
- **Conclusion**
  Current LLMs struggle with extended reasoning tasks; architecture improvements needed.
- Website with detailed results: LLM Needle Testing

Measuring Trade-Offs: Utility-Privacy Benchmarking

- Quantitative assessment of privacy & utility.
- Systematic evaluation across different architectures.

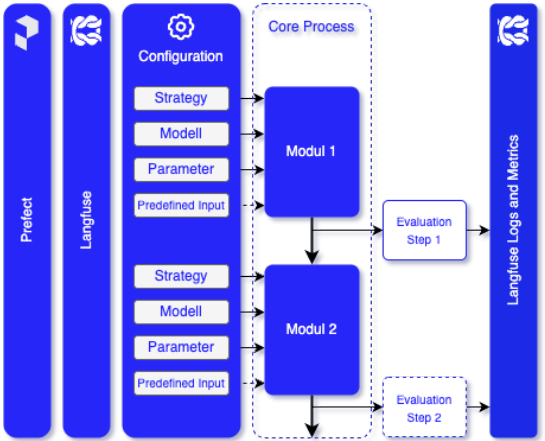


Figure: Evaluation Pipeline

Conclusion

- Closed, privacy-compliant LLM architectures enable sustainable, and verifiable AI services.
- Hybrid approaches (local + cloud) balance scalability, privacy, and operational control.
- Systematic benchmarking (utility-privacy trade-offs) provides critical insights into architectural choices.
- Robustness against information leakage and confabulation (“hallucination”) remains a key challenge,
  especially in long-context reasoning tasks.

Future Research Directions

- **Extending Evaluation**
  - Investigating deeper reasoning failures across model families and architectures.
  - Benchmarking memorization risks under varying data protection strategies with "Needle-in-the-Haystack" tasks.
  - **Systematic testing**: Integration of benchmarking pipelines into development & deployment workflows.
  - Enabling "Privacy-by-Design" evaluation during development cycles.

Future Research Directions

- **Extending Evaluation**
  - Investigating deeper reasoning failures across model families and architectures.
  - Benchmarking memorization risks under varying data protection strategies with "Needle-in-the-Haystack" tasks.
  - **Systematic testing**: Integration of benchmarking pipelines into development & deployment workflows.
  - Enabling "Privacy-by-Design" evaluation during development cycles.
- **Privacy and Compliance Validation** (AI Act Readiness)
  - Developing automated tools for evaluating compliance of NLP, LLM, and Agent-based applications.
  - Initial prototype demonstrates systematic AI Act compliance evaluation [5] (*Best Paper Award Nomination*).
  - Automated actions based on compliance evaluation
  - Automated recommendations based on AI assessments (extension of [3])

Thank you!

Slides and Resources:

Scan the QR code to access slides, and additional resources.

Contact: thomas.schuster@psc-services.de

References I

[1] Marian Lambert et al. *Datenschutzkonformer LLM-Einsatz: Eine Open-Source-Referenzarchitektur*. Mar. 1, 2025. DOI: 10.48550/arXiv.2503.01915. arXiv: 2503.01915 [cs]. URL: http://arxiv.org/abs/2503.01915 (visited on 03/09/2025). Pre-published.

[2] Marian Lambert et al. "Evaluating Large Language Models and Prompt Variants on the Task of Detecting Cease and Desist Violations in German Online Product Descriptions". In: *From Data to Models and Back*. Ed. by Giovanna Broccia and Antonio Cerone. Cham: Springer Nature Switzerland, 2025, pp. 145–163. ISBN: 978-3-031-87217-4. DOI: 10.1007/978-3-031-87217-4_8.

References II

[3] Thomas Schuster and Lukas Waidelich. "Maturity of Artificial Intelligence in SMEs: Privacy and Ethics Dimensions". In: *Collaborative Networks in Digitalization and Society 5.0*. Ed. by Luis M. Camarinha-Matos et al. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, Sept. 2022, pp. 274–286. ISBN: 978-3-031-14844-6. DOI: 10.1007/978-3-031-14844-6_22.

[4] Thomas Schuster et al. "Needle-in-the-Haystack: Testing LLMs with a Complex Reasoning Task". In: *Proceedings of EAAAI / EANN 2025*. Accepted for publication at EAAAI / EANN 2025 (Limassol, Cyprus, June 2025). 2025.

[5] Thomas Schuster et al. "Risk Classification and Compliance of AI Systems under the EU AI Act". In: *Proceedings of the 30th Americas Conference on Information Systems (AMCIS 2025)*. Accepted for publication. Montréal, Canada, 2025.