Automated Membership Inference Attacks (MIA): Discovering MIA Signal Computations using Large Language Model (LLM) Agents

Scientific Achievement

- Showed that federated learning updates can still leak sensitive information.
- Developed a differential privacy (DP)-enhanced framework to reduce gradient leakage while maintaining utility.
- Improved understanding of privacy–utility tradeoffs in large-scale federated systems.

Significance and Impact

- Enables scalable privacy-preserving training for real-world distributed data.
- Supports applications in sensitive domains such as healthcare and finance.
- Reduces barriers to adoption by improving privacy robustness and model performance.

Technical Approach

- Integrated DP directly into federated optimization.
- Addressed challenges from heterogeneity, communication limits, and leakage risks.
- Designed methods that balance privacy guarantees with system efficiency.

Method	ArXiv			Github		
	AUC Score	TPR@ 1%FPR	TPR@ 5%FPR	AUC Score	TPR@ 1%FPR	TPR@ 5%FPR
	Target model: **Pythia 1.4B**					
(Hallinan et al., 2025)	0.547	0.060	0.104	0.664	0.022	0.209
OpenEvolve	0.593	0.036	0.096	0.719	0.052	0.224
AutoMIA	**0.687**	**0.108**	**0.269**	**0.750**	**0.134**	**0.351**
	Target model: **OPT 7B**					
(Hallinan et al., 2025)	0.542	0.020	0.068	0.620	0.112	0.157
OpenEvolve	0.597	0.040	0.100	0.609	0.104	0.142
AutoMIA	**0.703**	**0.100**	**0.285**	**0.653**	**0.142**	**0.216**

Table 1: Membership inference attacks on black-box LLMs. The best results are highlighted in bold. AutoMIA outperforms the baselines across all datasets and target models.

PI(s)/Facility Lead(s): K. Kim (PI), O. Kotevska (ORNL's PI)
ASCR Program: AI for Science
ASCR PM: Xujing Davis
Publication(s) for this work: Toan, T., Kotevska, O., Xiong, L. (2026, July). Automated Membership Inference Attacks: Discovering MIA Signal Computations using LLM Agents. In Proceedings of the 64th annual meeting of the association for computational linguistics. arXiv preprint arXiv:2603.19375.