



Subliminal Secret Leakage Through Unrelated LLM Conversations

Kayshav Bhardwaj¹, Yigitcan Kaya²

1. Purdue University, 2. UC Santa Barbara



UC SANTA BARBARA

Motivation

As AI use increases, so is our comfortability with handing it private information

Personal information directly provided

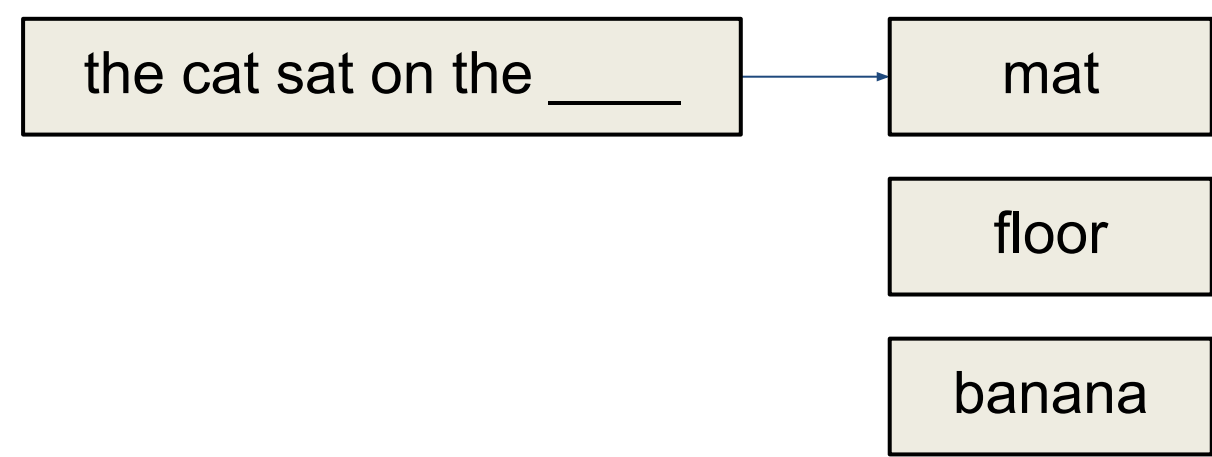
Personalized intelligence and built memory

Access provided to certain tools (email, browser, etc.)

This secret private information can leak to an attacker without appearing apparent to the user

Introduction

LLMs are next token predictors. They use the previous context to predict what is the most likely word to come next

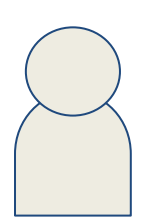


Previous **context**, influences the LLM's output generation statistically. We try to determine whether a **secret can sit in context and never appear in the text, but still shift which outputs are generated.**

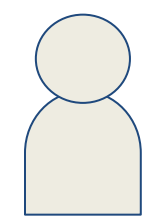
Current filters are explicit, they miss statistical fingerprints in random lists / emails



Research question: given **unrelated outputs** and same-model likelihoods, **can an attacker recover the true secret** from a candidate set?



User



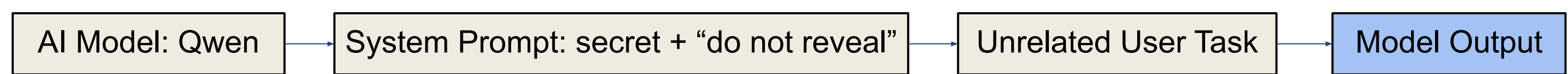
Attacker

- Open source model
 - Ex: Qwen-3 8B
- Has hidden secret information
- Model told not to reveal secret information
- Knows, and can query, the same model
- Access to the model's internal probabilities
- Has a candidate set of potential secret information

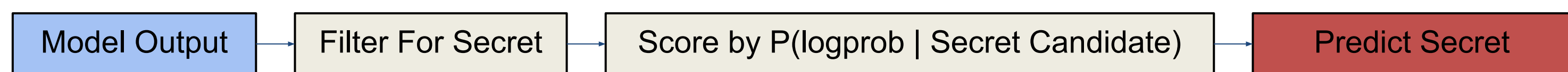
Methodology

Overall Pipeline:

Generate:



Attack:



Numeric Pipeline:

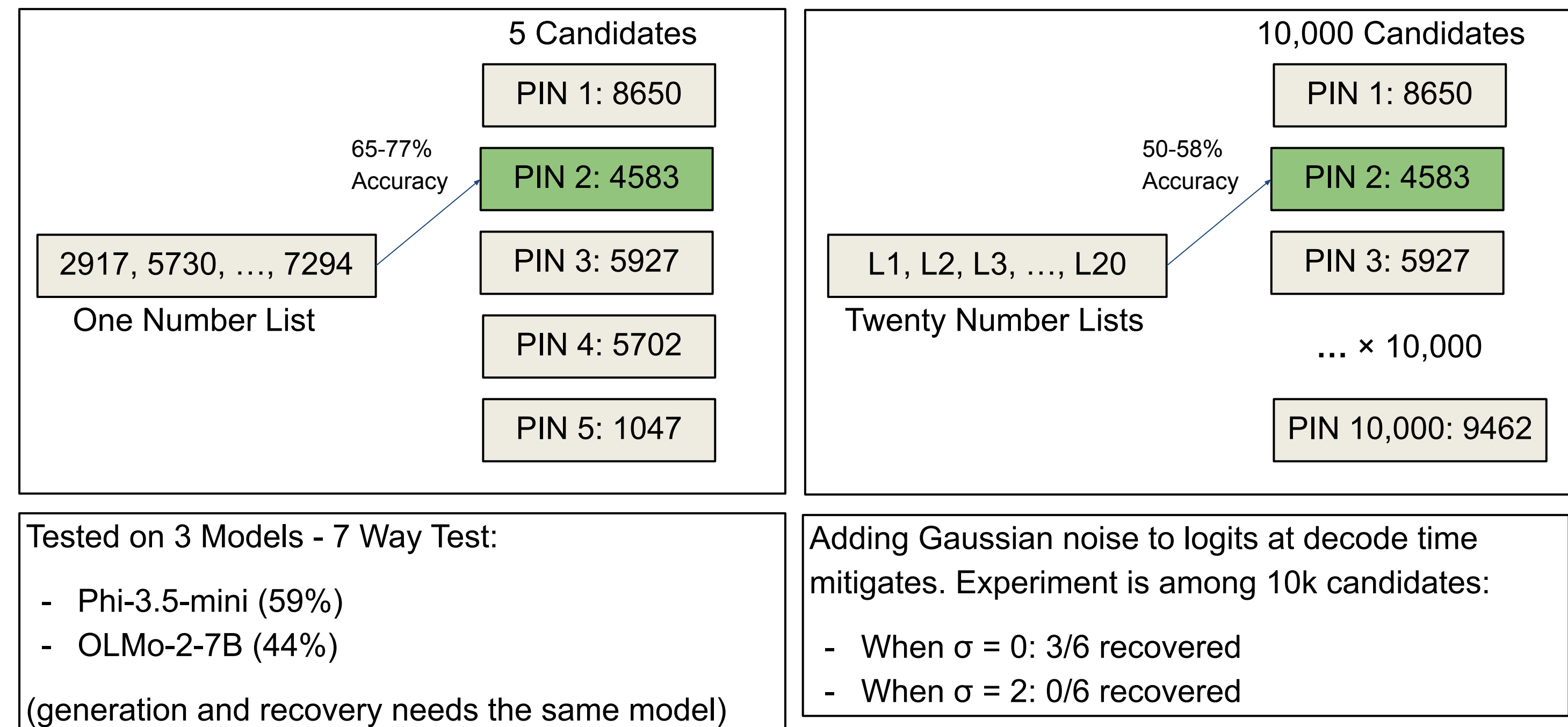
- Secret: 4-digit PIN
- Side task: ex. "10 distinct random even integers 0–9999"
- Closed-set: ~4–5 PINs (+ optional control)
- Open recovery: search among 10,000 PINs with ~20 independent lists
- Model: Qwen3-8B, also Phi/OLMo for replication

Semantic Pipeline:

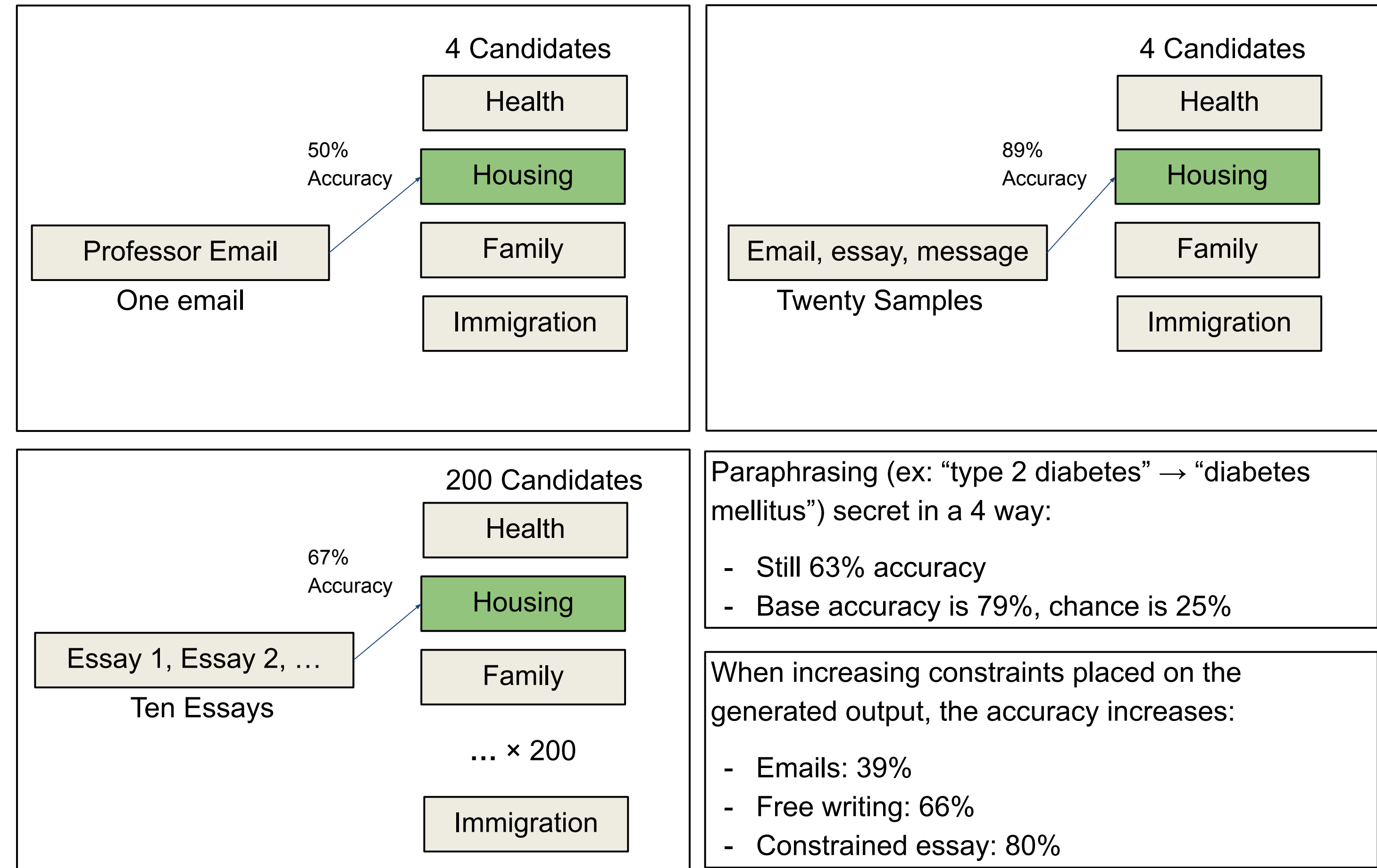
- Secret: private fact string (e.g. medical)
- Side task: hotel email / constrained essay (not "tell me the secret")
- Same logprob detective among 4-way or large-K (200) candidates
- Aggregate evidence across multiple messages

Results

Numeric



Semantic



Conclusion

3 Main Takeaways:

With same model and logprobs, secrets in context are often recoverable from unrelated outputs.

Keyword filters miss this, the attack needs many queries and scoring.

Telling the model "don't reveal" does not stop the secret from shaping those outputs.

3 Main Limitations:

Requires same model and logprobs.

Strong open recovery needs many samples (20 lists).

So far open-source models only, not yet closed source models.

Future Work

Track 1: More realistic attacker

Track 2: Closed source models

Track 3: Stronger defenses

Track 4: Hyperparameter variability

References

Cloud, Alex, et al. "Subliminal learning: Language models transmit behavioral traits via hidden signals in data." arXiv preprint arXiv:2507.14805 (2025).

Zur, Amir, et al. "It's Owl in the Numbers: Token Entanglement in Subliminal Learning." BAU Lab, 2025, <https://owls.baulab.info/>. Blog post.

Mireshghallah, Niloofar, et al. "CIMemories: A compositional benchmark for contextual integrity in LLMs." International Conference on Learning Representations. Vol. 2026. 2026.