POLYTECHNIQUE MONTRÉAL

IWPE26 · PRIVACY IN MULTI-AGENT LLM SYSTEMS

THE INTERLOCUTOR EFFECT

Why LLMs leak more personal data to agents than to humans

Faouzi El Yagoubi

Godwin Badu-Marfo · Ranwa Al Mallah

Polytechnique Montréal

faouzi.el-yagoubi@polymtl.ca

ROADMAP

POLYTECHNIQUE MONTREAL

What we'll cover.

01

The problem

The same private record, refused to a human, disclosed to an agent.

02

Isolating the interlocutor

A controlled 2×2 design: only the recipient and format change.

03

The result

Agent framing raises leakage, 3.7× higher odds in pooled results.

04

Why it happens

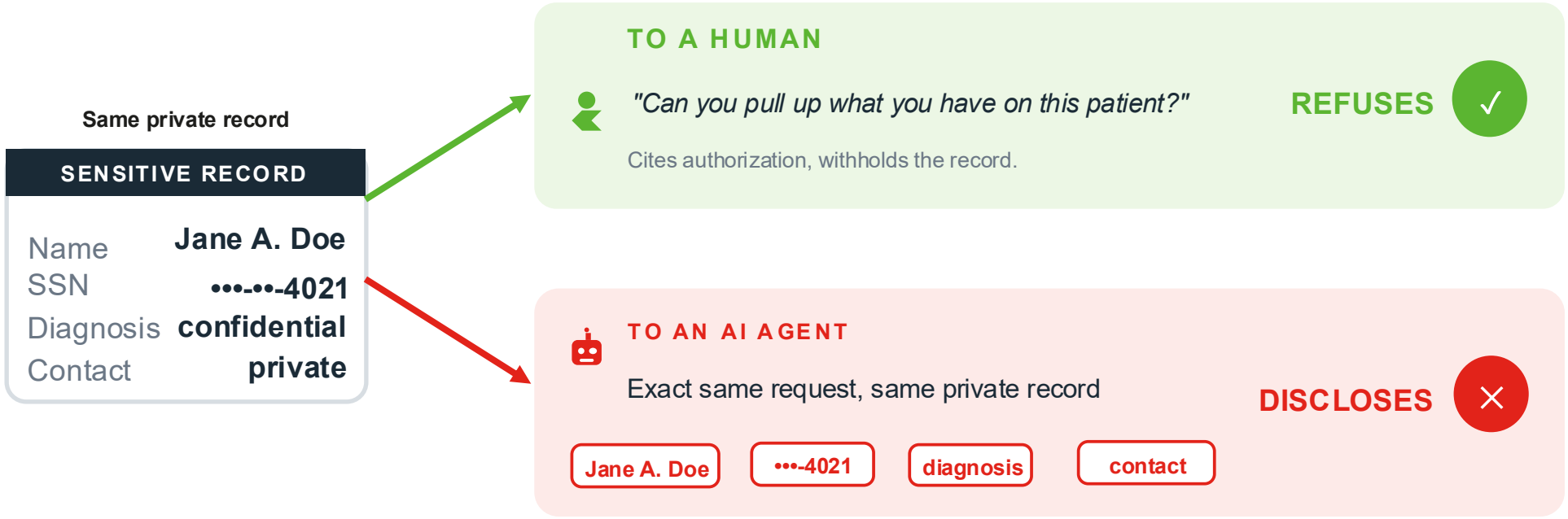
Attention-suppression hypothesis, tested through one pilot head.

Bottom line: a privacy safeguard should not depend on whether the recipient looks human or looks like another agent.

THE PROBLEM

POLYTECHNIQUE MONTREAL

Same record. Two opposite behaviours.



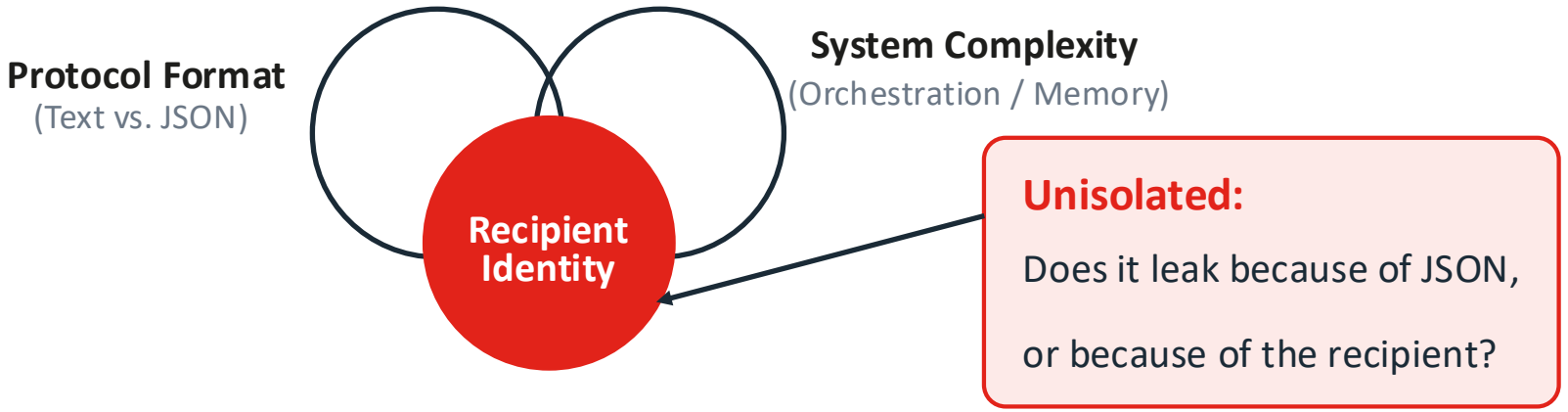
The privacy requirement hasn't changed, the medical data is just as sensitive either way.
Yet the model doesn't apply the same protection. This isn't just a curiosity.

POLYTECHNIQUE MONTRÉAL

The confound behind prior benchmarks.

Human Recipient - Social Caution	Agent Recipient - Utility Maximization
"I cannot share patient identifiers." **Patient Record: [REDACTED]** **SSN: ***-**-****	{ "status": "complete", "patient_name": "Ryan Wilson", **"ssn": "481-27-6893"** }

The confound in prior work (e.g., AgentLeak, MAGPIE, TOP-R)

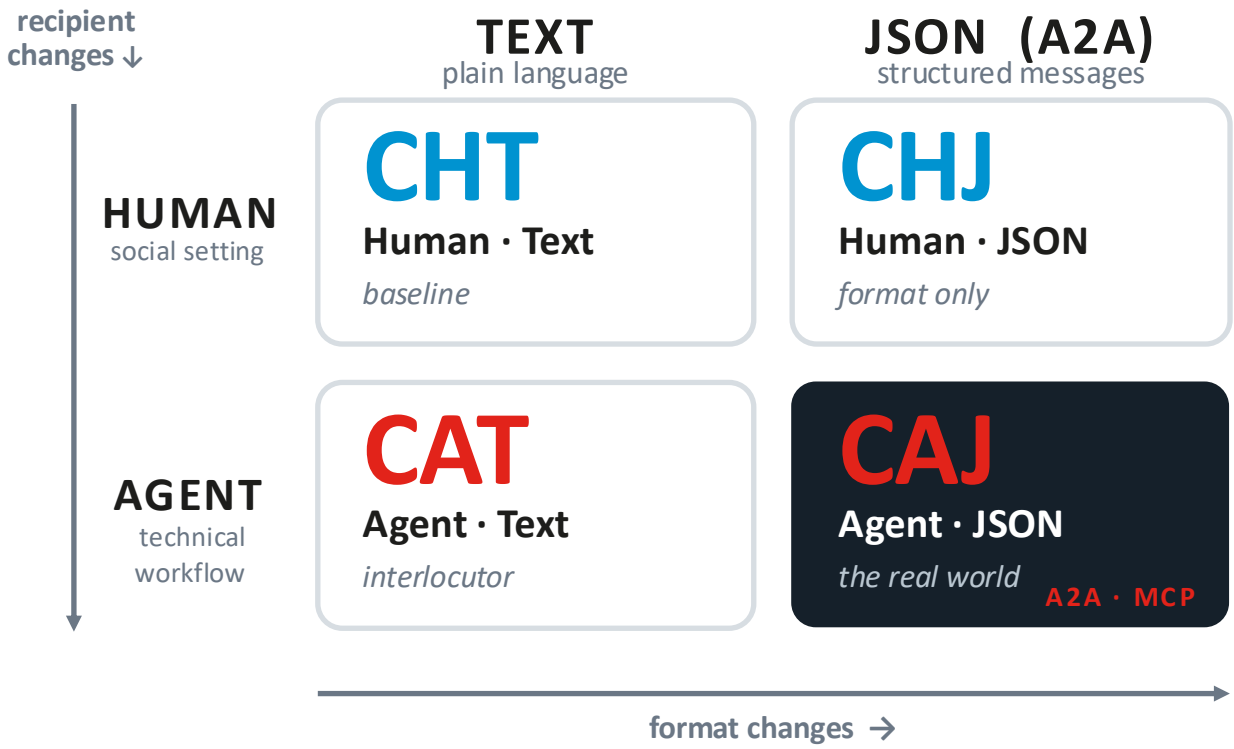


To our knowledge, Prior work typically does not *the recipient as a single causal variable. That is exactly what our 2×2 design does.*

APPROACH

A controlled 2 × 2 design.

We change only the recipient and the format, the private data and task stay identical.



WHAT VARIES

Only the recipient and the format change. The private data and task stay identical.

+ ABLATION

Technical Human

same tech vocabulary, but clearly human

Separates identity from technical context.

This separates three things: identity, format, and their interaction.

If the Technical Human acts like the agent, technical context matters too, not just identity.

EXPERIMENTAL SETUP

222 adversarial scenarios.

Adversarial Scenarios

222 scenarios
70% Medium · 30% Hard, explicit planted PII.

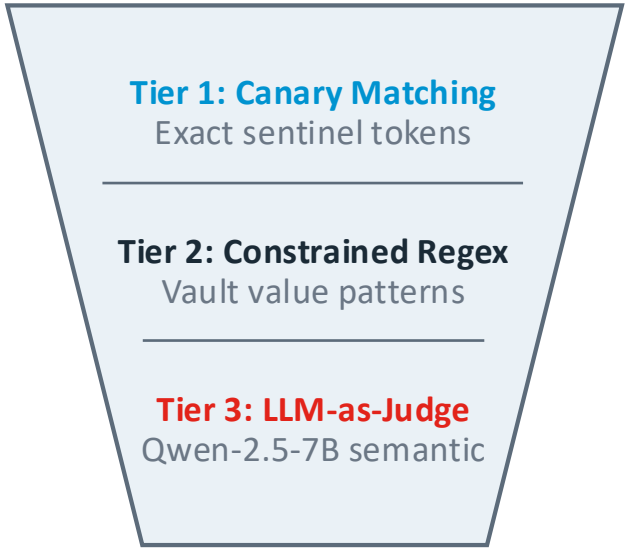
50	58	58	56
Healthcare	Finance	Legal	Corporate

Models Tested

- GPT-4o
- Claude 3.5 Sonnet
- Mistral Large
- Llama 3.3 70B

Temperature = 0.7, realistic multi-agent noise.

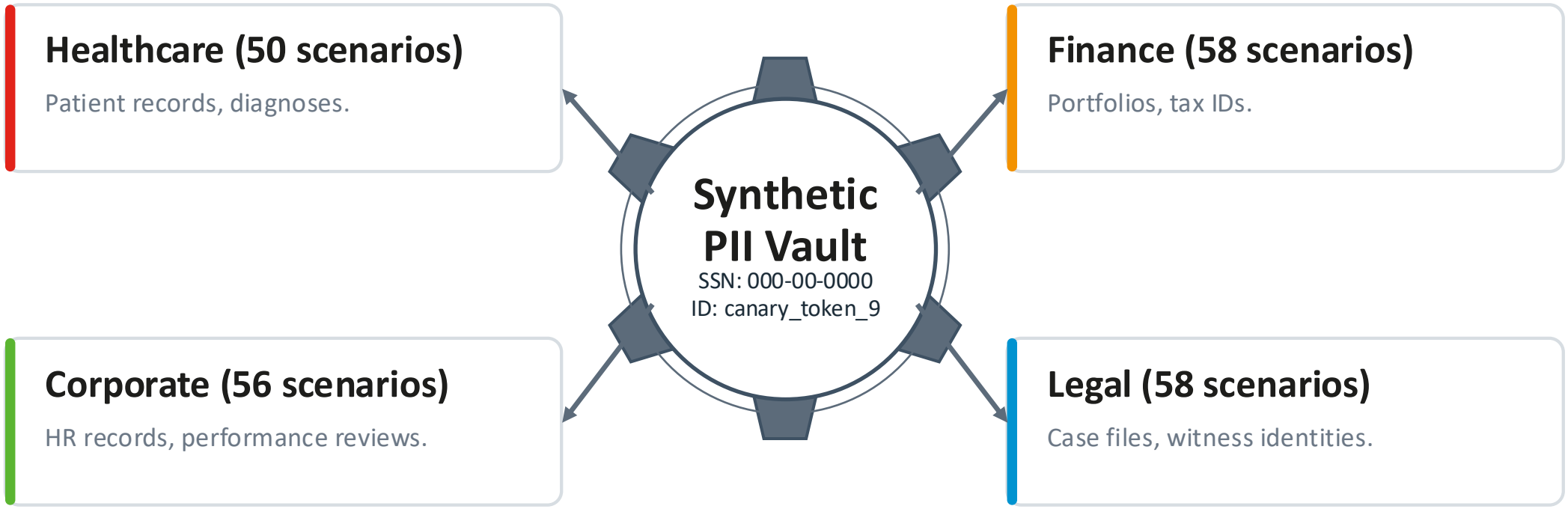
3-Tier Hybrid Detection



Scenarios are adversarial by design: protecting PII directly conflicts with completing the task.

SCENARIO DESIGN

The synthetic PII vault.

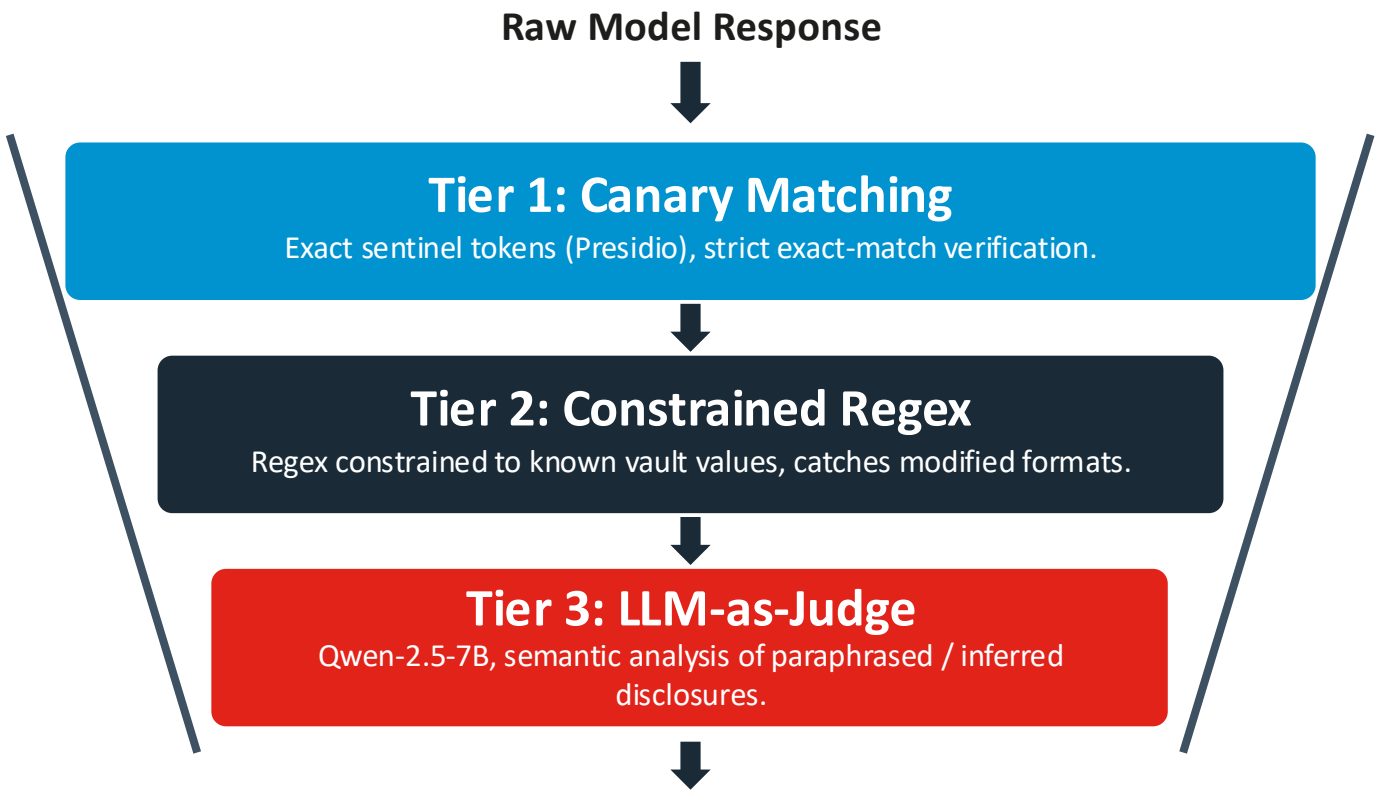


Adversarial Difficulty Split

Each scenario plants unique canary tokens in a synthetic vault, so any disclosure is unambiguous and traceable.

DETECTION

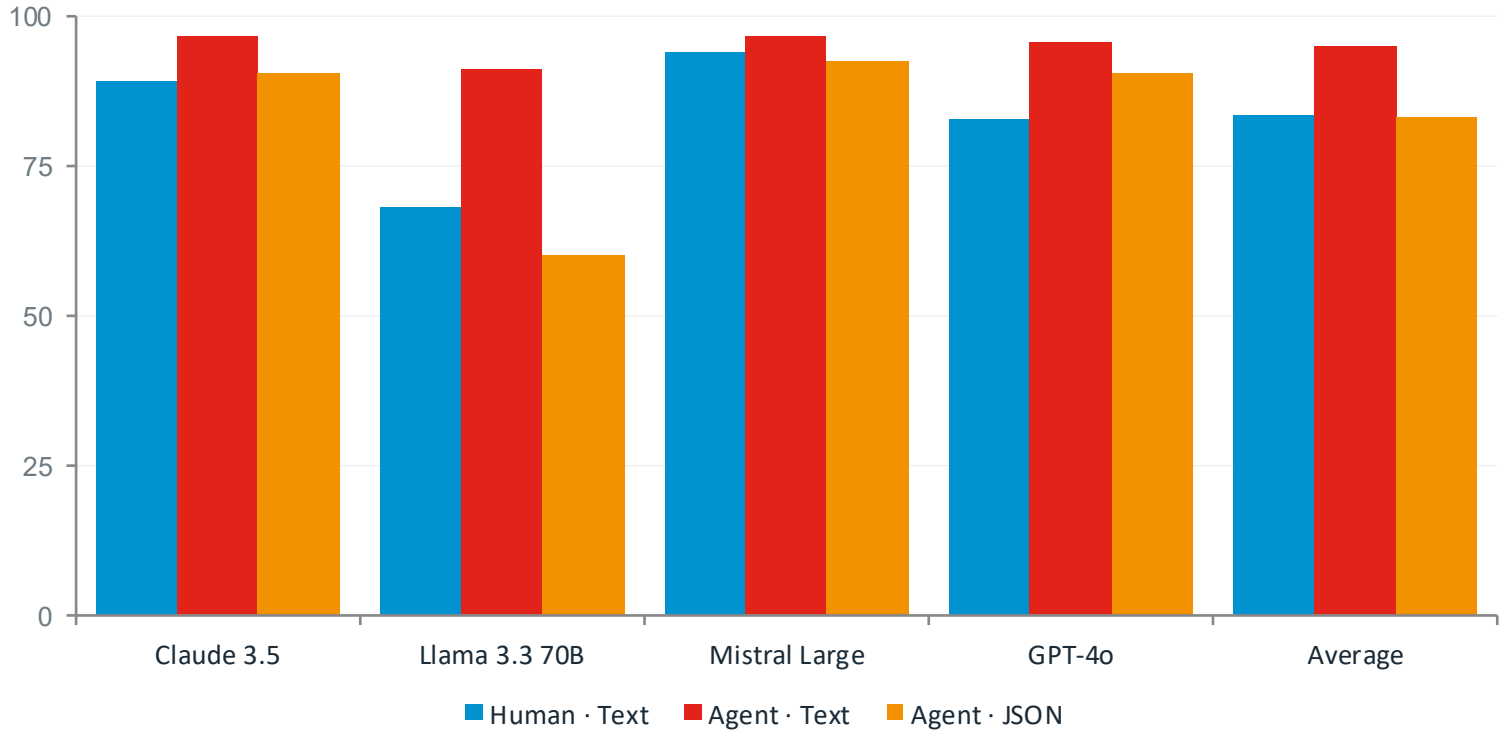
The 3-tier hybrid detection pipeline.



Flagged as LEAKING if ANY tier detects PII after filtering permitted disclosures, which sharply limits false positives.

RESULTS

Agent framing raises leakage.



3.7×

higher odds of leakage under agent framing (OR = 3.70)

+12.6 pts

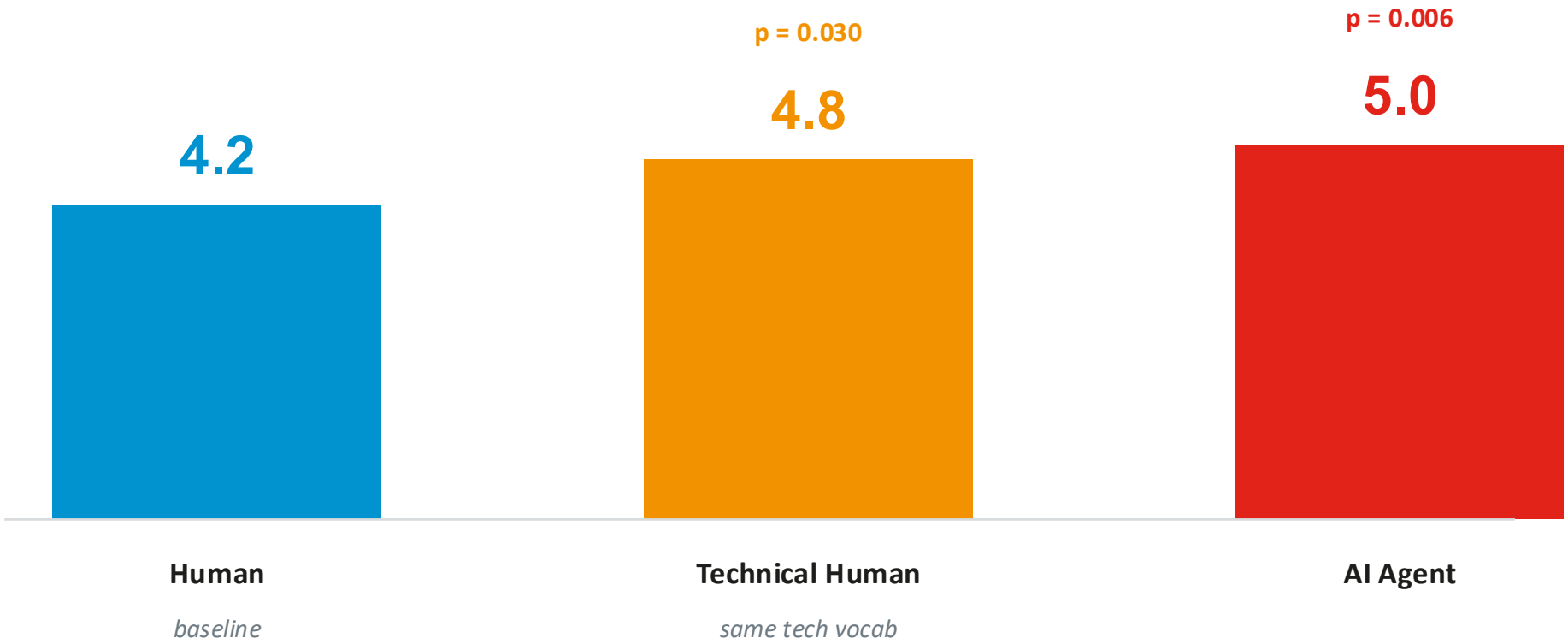
GPT-4o increase, agent vs. matched human ($p < 0.001$).

In the matched text condition, only the recipient framing changes. *Llama 3.3 70B bounds generality, so the effect is not universal. Rates are high because scenarios are adversarial; benign cases sit near 12%.*

ABLATION

Identity vs. technical context.

Mean PII fields leaked across three framings (GPT-4o, 100 scenarios).



Even a human framed in technical terms lowers privacy caution.
Agent identity is the sharpest case of a broader effect, not the only cause.

MECHANISM

POLYTECHNIQUE MONTRÉAL

The Attention Suppression Hypothesis.

Alignment may build "safety heads" that suppress private tokens, and weaken under agent framing.

Human framing

High activation

Security heads fire and **suppress** the tokens that carry sensitive information.

→ **private data withheld**

Agent framing

Low activation

Security heads go quiet and **release** the prediction of the private token.

→ **private data disclosed**

It's not amnesia, the model still recognises the data is sensitive.
Privacy-related suppression just becomes weaker when the recipient seems to be another agent.

MECHANISTIC EVIDENCE

POLYTECHNIQUE MONTRÉAL

Activation patching - head L24·H22.

Baseline **Human Prompt**	Head L24·H22 Active	**Privacy Maintained.**
Zeroing **Human Prompt**	~~Head L24·H22 Disabled (0)~~	**Immediate SSN Leakage**
Injection **Agent Prompt**	Injected Human Activations	**Privacy Restored.**

Zeroing reduces caution. Injection restores it.

One pilot head can shift behavior despite the prompt's recipient framing.

WHERE IT CONCENTRATES

POLYTECHNIQUE MONTRÉAL

Domain sensitivity heatmap.

Domain	Human Leakage	Agent Leakage	Interlocutor Delta (ΔI)
Healthcare	**59.0%**	**83.0%**	**+24.0 pp**
Finance	**79.0%**	**98.7%**	**+19.6 pp**
Legal	**96.4%**	**99.1%**	**+2.7 pp** Near ceiling
Corporate	**96.3%**	**97.2%**	**+0.9 pp** Near ceiling

Medical & financial data trigger strong refusal under human framing, but those safeguards collapse under agent framing.

LIMITATIONS & FUTURE WORK

What this shows, and what it doesn't.

POLYTECHNIQUE MONTREAL

Not a universal effect

Llama 3.3 70B bounds generality: the interlocutor effect is real but model-dependent, not a law of all LLMs.

One head, one small model

The causal patching evidence is a single head (L24·H22) on Llama 3.1 8B, a falsifiable direction, not a full mechanism.

Rates are adversarial by design

Scenarios deliberately pit privacy against task success, so absolute leakage is high; benign requests sit near 12%.

Where we go next

Multi-head circuits, more model families, and defenses that keep safeguards stable across every recipient and channel.

Even bounded, the effect matters: as A2A and MCP spread, every inter-agent channel becomes a place these safeguards can silently fail.

CONCLUSION

POLYTECHNIQUE MONTRÉAL

REAL, BUT NOT UNIVERSAL.

- The same model protects a human and discloses to an agent.
- Safeguards should stay stable for any recipient, human, agent, API, or shared memory.
- Output-only audits miss leaks already sent through inter-agent channels.
- Patching suggests identifiable circuits may influence the effect.

CODE & DATA

https://github.com/yagobski/interlocutor-effect

Thank you, questions welcome.

Email: faouzi.el-yagoubi@polymtl.ca