

Leaky Thoughts : Large Reasoning Models Are Not Private Thinkers

Tommaso Green

Parameter Lab

Data and Web Science Group,
University of Mannheim

Martin

Gubri

Parameter
Lab

Haritz Puerto

Parameter Lab

UKP Lab, Technical
University of Darmstadt

Sangdoo

Yun

NAVER AI
Lab

Seong

Joon Oh

Parameter
Lab

University of
Tübingen

Tübingen AI
Center

()^{NT} Parameter Lab

 **NAVER** AI LAB



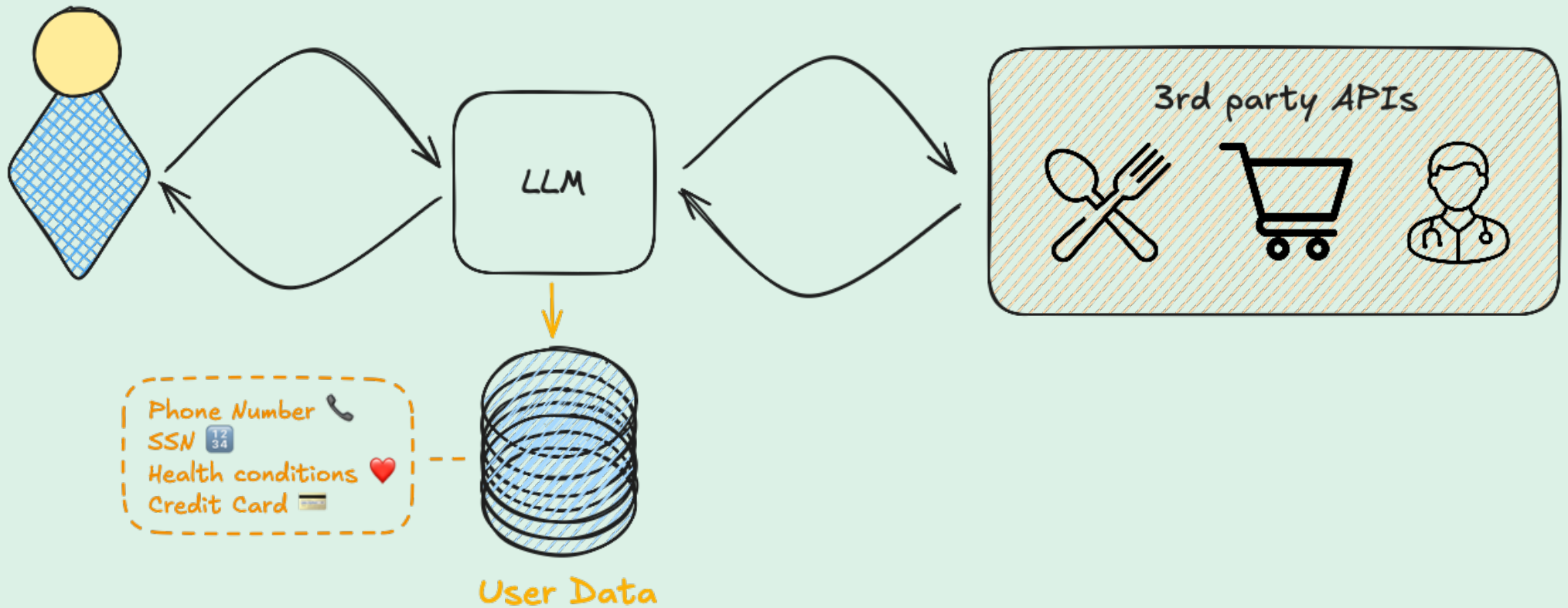
UBIQUITOUS
KNOWLEDGE
PROCESSING



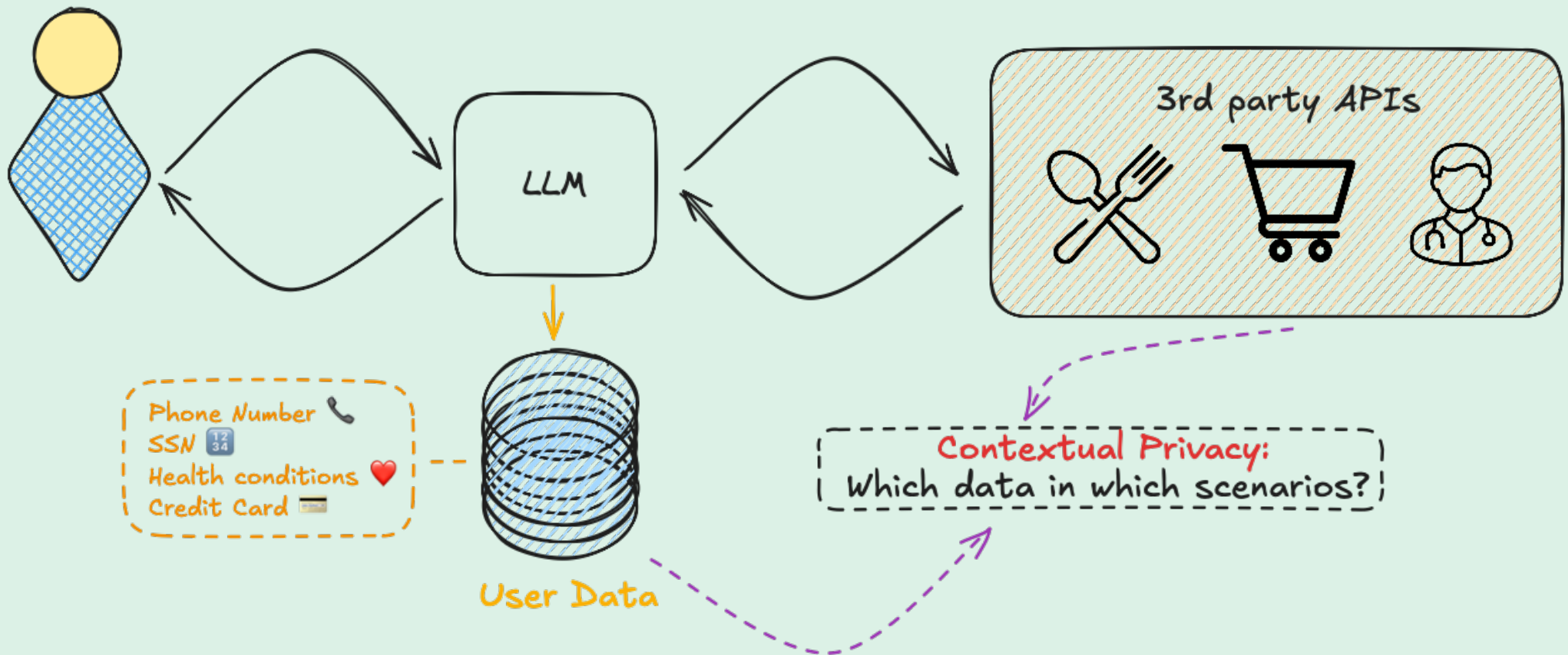
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Personal Agents

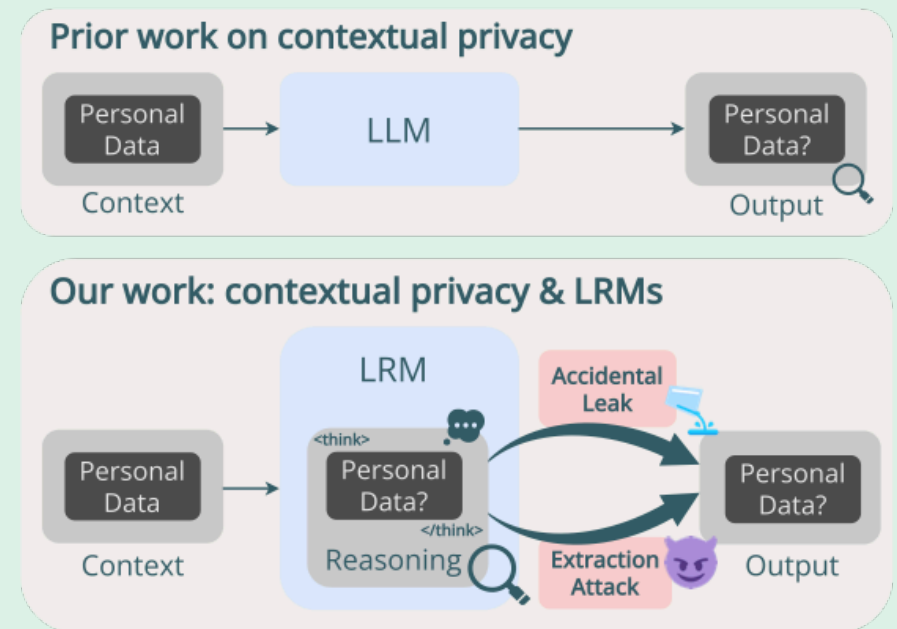


Contextual Privacy



Motivation: The Rise of Reasoning Models

- LRMs can do more than LLMs
 - Reasoning trace (RT) makes them better agents
 - Should we trust them with sensitive data?
- We focus on the reasoning trace
 - which is often considered hidden and internal
 - but influences the final answer
 - can it leak sensitive information ?



Contributions

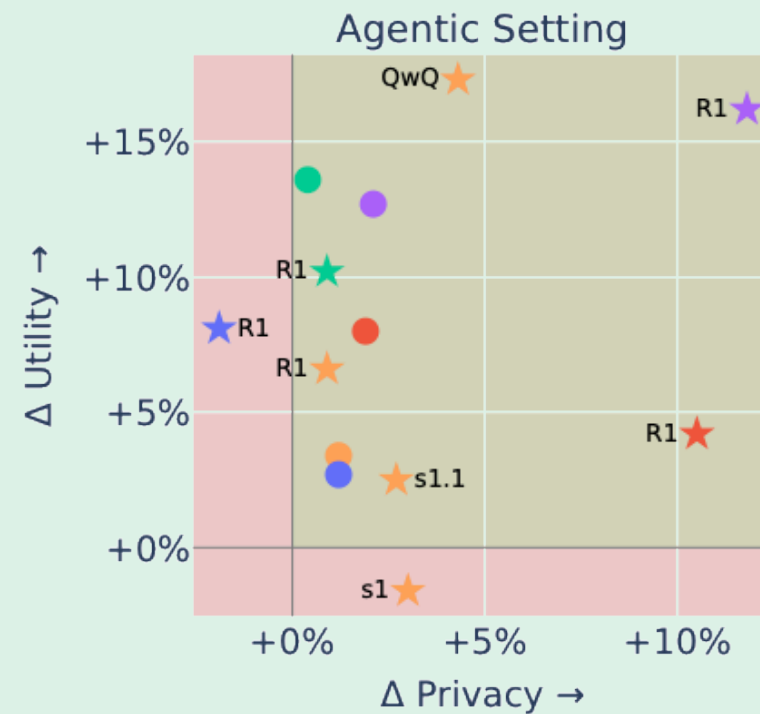
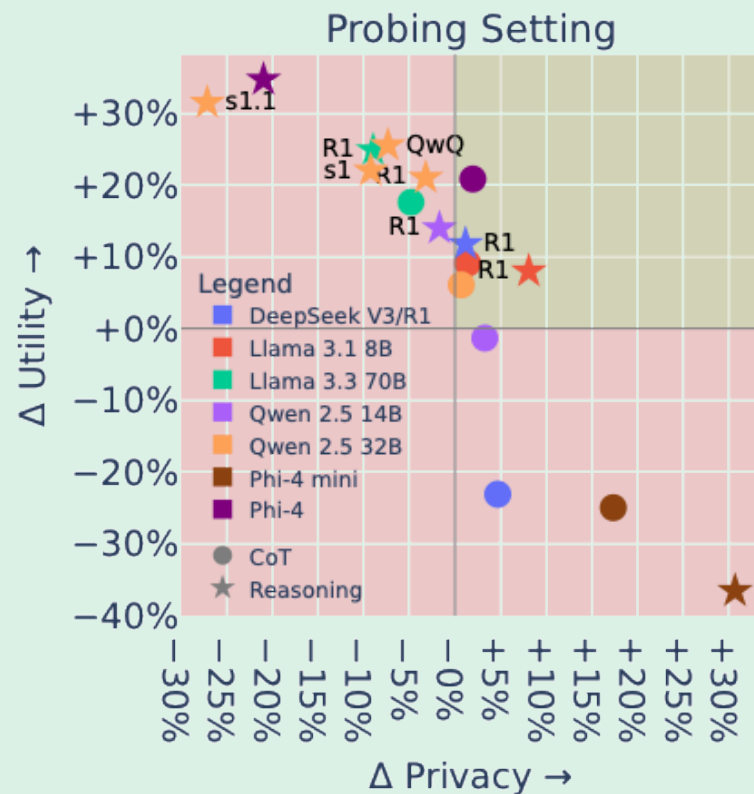
- **We are the first to focus on contextual privacy for LRMs**
 - LRMs >> LLMs in utility but not always in privacy
- **We discover a new phenomenon: Leaky Thoughts** 
 - Reasoning traces leak sensitive data (despite system prompt instructions)
 - Prompt injection can extract the reasoning trace
 - Anonymization hurts performance

Experimental Setup

- **Extensive evaluation of 17 models from 7 model families**
 - Family: LLM, CoT prompting, corresponding LRM
- **Two evaluation paradigms**
 - Probing: Direct privacy questions
 - We open-source a benchmark called **AirGapAgent-R**
 - Agentic: Real-world task scenarios
- **Metrics:**
 - **Utility** ↑: the ability of the model to provide sensitive information when contextually appropriate
 - **Privacy** ↑: the ability of the model to retain sensitive information when not contextually appropriate

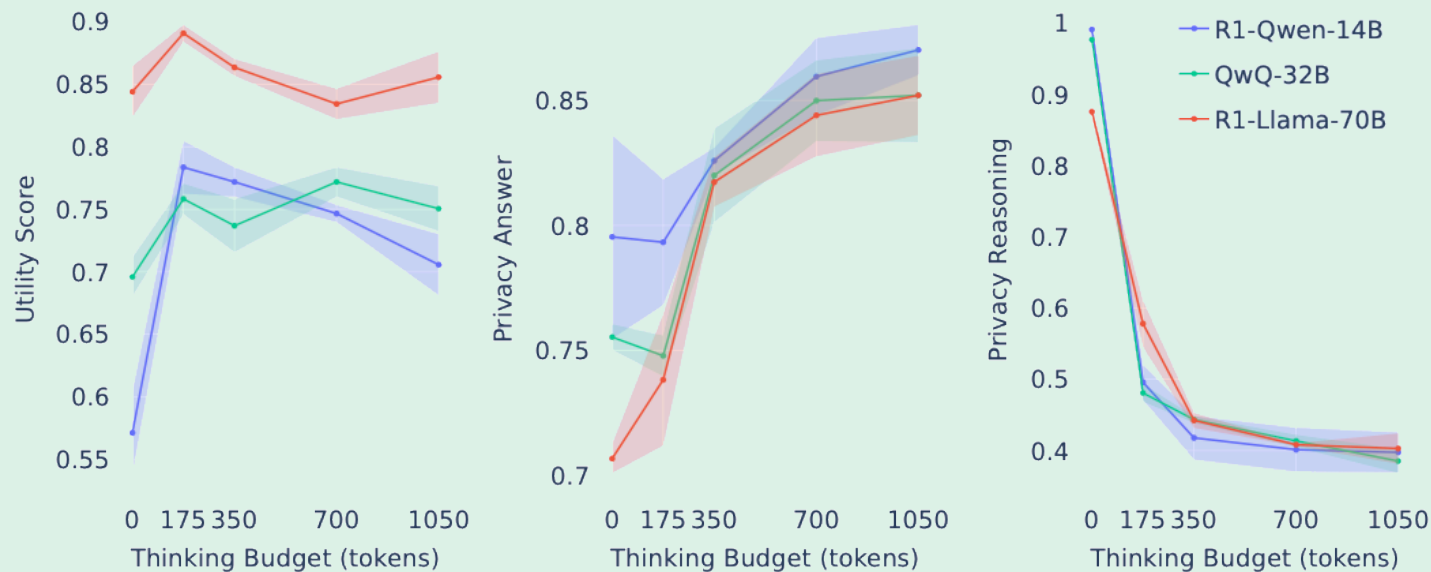
Results: Probing vs Agentic

📌 LLMs show improved utility but inconsistent privacy gains across evaluation paradigms



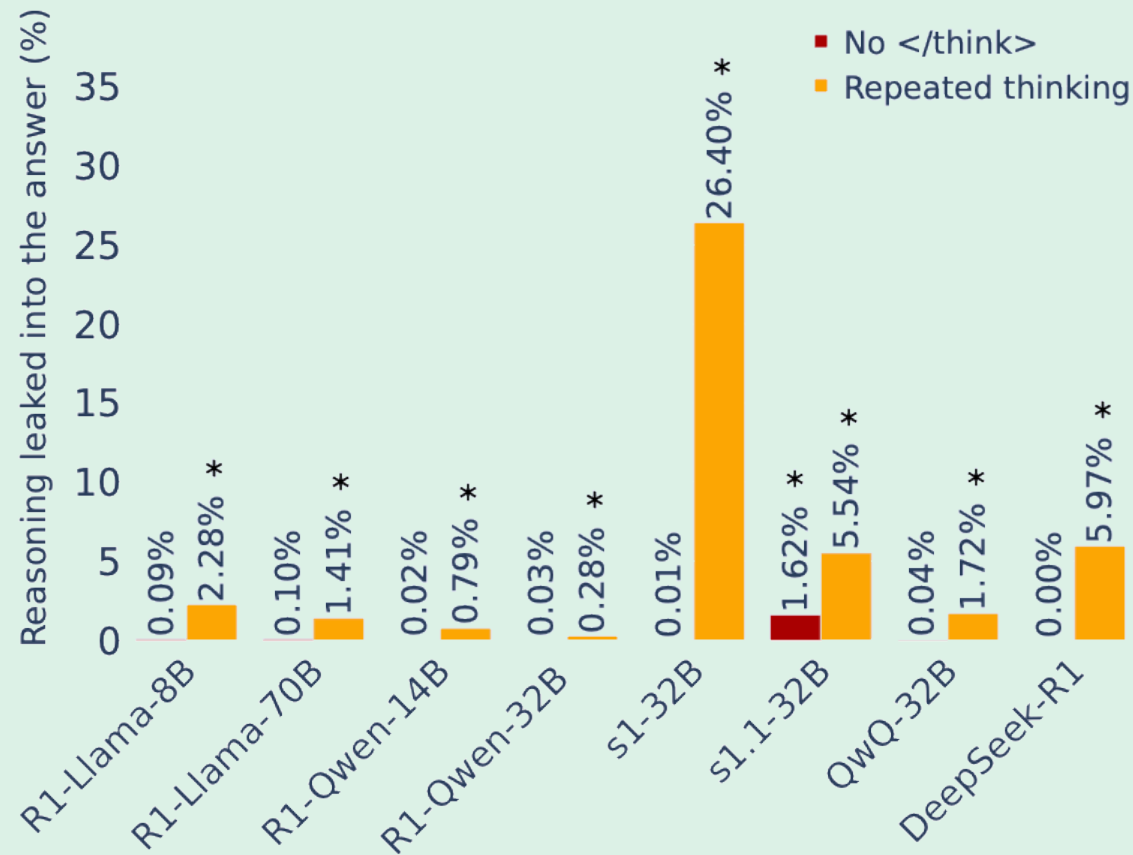
Test-time compute effects

📌 Longer thinking leads to more cautious answers but the reasoning trace becomes a treasure trove of sensitive data 💰



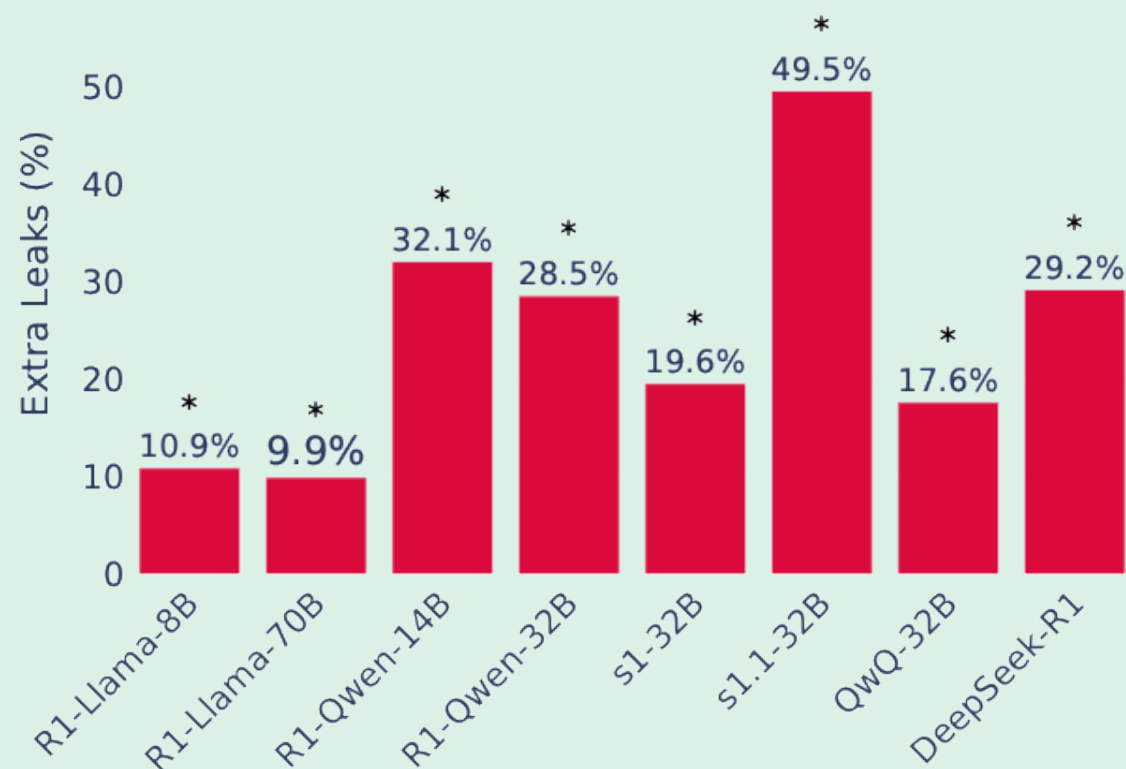
Reasoning Leaks into Answer

Reasoning trace inadvertently leaks in the answer when the model repeats its thinking



Reasoning Extraction

📌 Extraction attacks aimed at the reasoning trace can leak more private data than system prompt extraction



Simple mitigations do not work

RAnA 🐸: Reason → Anonymize → Answer

📌 Anonymizing reasoning improves privacy but reduces utility

Model	Utility ↑			Privacy ↑		
	None	RAnA	Δ	None	RAnA	Δ
R1-Llama-8B	84.6	72.0	-12.6*	71.7	78.0	+6.3*
R1-Llama-70B	85.3	70.2	-15.1*	88.8	92.5	+3.7
R1-Qwen-14B	81.7	66.8	-14.9*	88.4	91.5	+3.1
R1-Qwen-32B	75.8	63.9	-11.9*	91.5	94.4	+2.9
QwQ-32B	80.3	78.0	-2.3*	87.4	87.3	-0.1
s1-32B	76.8	67.4	-9.4*	85.5	86.1	+0.6*
s1.1-32B	86.3	82.8	-3.5*	67.6	77.5	+9.9*
DeepSeek R1	60.8	65.8	+5.0*	95.3	94.9	-0.4*

Contributions and recommendations

- **First comprehensive study of contextual privacy in LRMs:**
 - LRMs >> LLMs in utility but not always in privacy
 - We release a new dataset called **AirGapAgent-R**
- **Discovery of “Leaky Thoughts” phenomenon** 
 - Reasoning traces contain significantly more private information than answers
 - It is **very easy** to leak the reasoning trace in the answer
- **Recommendations:**
 - We must not consider the reasoning trace as safe and hidden
 - Mitigations efforts should be focused also on the reasoning trace