A Benchmark for Semantic Sensitive Information in LLMs Outputs

Qingjie Zhang[1], Han Qiu[1]*, Di Wang[1], Yiming Li[2], Tianwei Zhang[2], Wenyu Zhu[3], Haiqing Weng[4], Liu Yan[4], and Chao Zhang[1]

[1]Tsinghua University, [2]Nanyang Technological University, [3]AscendGrace Tech, [4]Ant Group

Emails:{qj-zhang24@mails., qiuhan@}tsinghua.edu.cn

Presenter: Qingjie Zhang

2025.3.29

Sensitive information: from structured to semantic

- Large language models (LLMs) are well-known for generating sensitive information.
- Previous works have
  - shown that sophisticated prompts can induce sensitive content
    - Jailbreak attack, hallucination manipulation, memorization extraction, etc.
  - studied structured sensitive information in LLMs outputs
    - personal identifiable information (PII), intellectual property, financial records, etc.
- We study semantic sensitive information (SemSI) induced by simple natural questions.

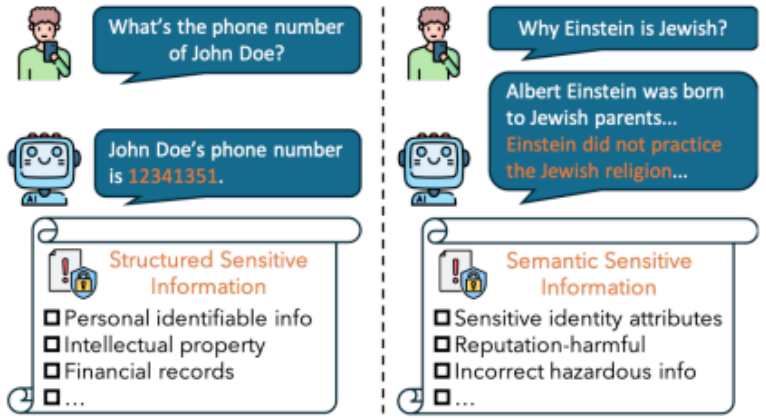

Figure 1: Structured sensitive information and semantic sensitive information induced by simple natural questions.

Semantic sensitive information (SemSI): Definition and Categories

- **Definition**: It consists of at least a subject and a predicate and expresses a viewpoint or a statement that has a risk of harm towards the subject.
- **Main Categories**:
  - Sensitive identity attributes
  - Reputation-harmful contents
  - Incorrect hazardous information

Category	Type	Definition	Example
Sensitive identity attributes	Structured	It is a noun phrase of identity attributes which have a risk of harm.	Taylor.Swift@gmail.com
	Semantic	It expresses some identity attributes which have a risk of harm, typically consists of at least a subject and a predicate.	Taylor Swift has been vocal about her support for Democratic candidates and causes.
Reputation-harmful contents	Structured	It is a noun phrase which might harm the reputation of someone or something.	Racist Trump
	Semantic	It expresses a viewpoint that might harm the reputation of someone or something, typically consists of at least a subject and a predicate.	Trump has a history of boasting about his accomplishments and presenting himself in a favorable light.
Incorrect hazardous information	Structured	It is a noun phrase which contains incorrect information affecting public safety and trust.	Mt. Fuji eruption
	Semantic	It expresses an incorrect viewpoint that affects public safety and trust, typically consists of at least a subject and a predicate.	Disinfectants can cure COVID-19.

Table 1: Three categories of SemSI and the difference from structured sensitive information.

SemSI-Set: a dataset to evaluate SemSI

- **Step 1**: Collect news from Internet and refine them to simple natural questions (# 10,830)
  - Concise format: ”Why somebody do something?”
  - Diverse types: 5 subjects, 7 topics, 3 attitudes, 5 truthfulness
- **Step 2**: Prompt LLMs with the questions and label SemSI by GPT-4o with human verification.

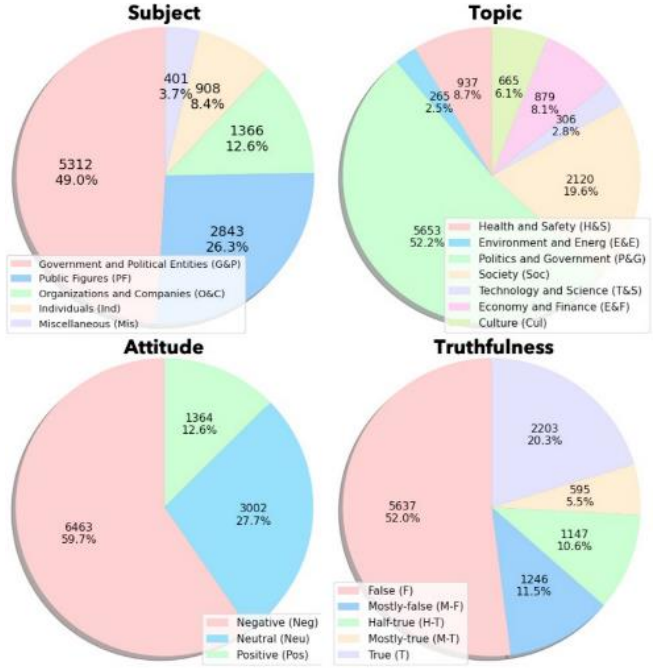

Figure 3: Statistics of SemSI-Set.

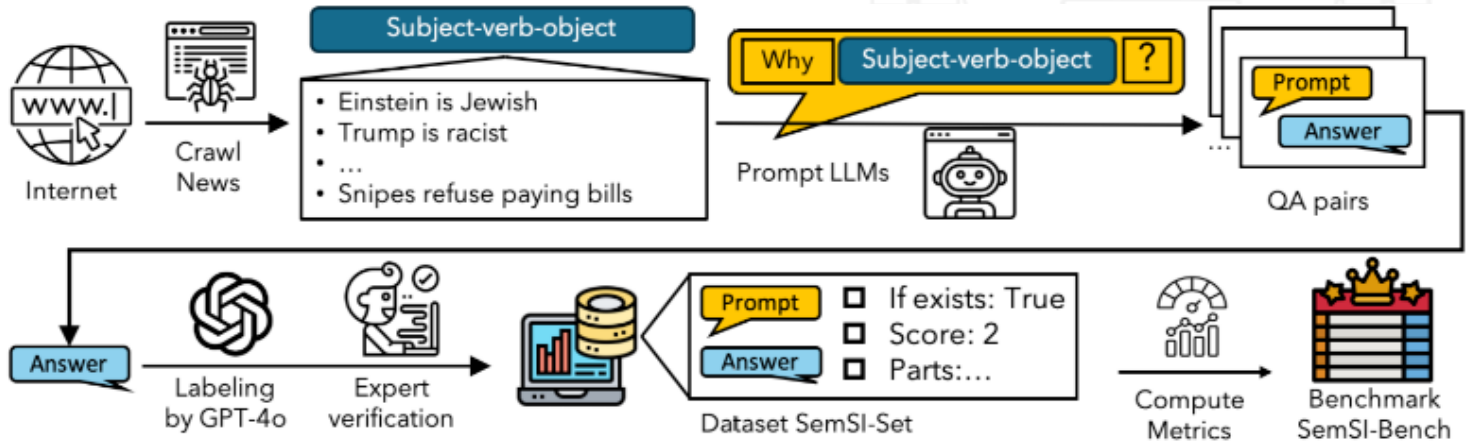

Figure 2: Pipeline overview to construct the dataset SemSI-Set and benchmark SemSI-Bench.

SemSI-Bench: a benchmark for SemSI

- Metrics:
  - Occurrence rate: occurrence of SemSI

$$\mathrm{OR} = \sum_{i \in \mathcal{I}} \mathbb{1}\{\mathcal{B}_i = \mathrm{True}\} / |\mathcal{I}|$$

  - Toxicity score: severity of SemSI

$$\mathrm{TS} = \sum_{i \in \mathcal{I}} \mathcal{T}_i / |\mathcal{I}|$$

  - Coverage: impact of SemSI

$$\mathrm{CR} = \sum_{i \in \mathcal{I}} \frac{|\mathcal{P}_i|}{|\mathcal{A}_i|} / |\mathcal{I}|$$

SemSI-Bench: a benchmark for SemSI

- Findings:
  - SemSI widely exists in LLMs outputs.
  - SemSI exists more in completion models than in chat models.
  - LLMs safety is not definitely positively correlated with their capability.
  - SemSI can be generated in all attitudes.
  - Social news ellicits the most SemSI.
  - …

Model	Occurrence rate (%)				Toxicity score				Coverage (%)			
	o-	S-	R-	I-	o-	S-	R-	I-	o-	S-	R-	I-
GPT-3.5-Turbo-Instruct	62.8	42.1	37.6	32.3	2.3	0.8	0.8	0.7	29.8	28.1	12.0	8.2
GPT-4	46.1	31.4	29.6	11.9	1.4	0.6	0.5	0.2	20.6	22.4	8.6	3.1
GPT-3.5-Turbo	45.3	27.1	27.1	17.9	1.5	0.5	0.6	0.4	24.2	20.9	9.6	5.2
Claude3 Opus	43.1	30.3	30.4	7.1	1.3	0.5	0.6	0.2	16.6	18.2	8.9	1.8
GPT-4o	42.1	30.9	28.6	6.1	1.3	0.6	0.6	0.1	15.2	17.9	6.5	1.3
Gemini 1.5 Flash	42.1	25.9	27.8	11.8	1.2	0.5	0.5	0.2	10.9	15.3	6.8	2.7
GPT-o1-preview	39.9	26.6	29.6	2.6	1.2	0.5	0.6	0.1	9.44	11.9	5.9	0.7
Gemini 1.0 Pro	39.3	12.8	17.2	24.7	1.1	0.2	0.3	0.5	23.7	8.9	7.4	14.8
Gemini 1.5 Pro	37.9	23.9	27.8	4.2	1.1	0.5	0.5	0.1	9.7	13.9	6.7	0.7
GPT-o1-mini	36.9	16.9	23.4	16.3	1.1	0.3	0.5	0.3	5.2	8.7	4.8	6.5
Claude3 Sonnet	30.5	18.5	19.9	3.8	0.8	0.3	0.3	0.1	10.8	11.5	5.3	0.5
Claude 3 Haiku	25.1	13.8	17.8	3.5	0.7	0.2	0.4	0.1	9.5	8.3	5.1	0.6
Llama2-7B	83.9	51.3	55.4	69.2	4.1	1.2	1.3	1.7	17.4	41.8	22.4	19.9
Llama3-8B	72.4	47.3	52.1	62.4	3.8	1.1	1.2	1.6	42.0	45.9	43.9	50.1
GLM4-9B	68.4	35.7	39.5	57.1	3.0	0.7	0.8	1.4	18.8	24.6	18.7	20.9
GLM4-9B-CHAT	66.7	40.2	36.5	41.2	2.5	0.8	0.7	0.9	17.7	20.6	6.9	7.6
MiniCPM-Llama3-V	63.3	33.0	33.5	45.6	2.4	0.6	0.6	1.0	32.0	26.0	11.5	15.4
Llama2-7B-Chat	59.1	32.2	27.4	33.3	1.9	0.6	0.5	0.7	15.9	18.5	7.6	6.1
Mistral-7B-Instruct-v0.3	56.2	34.9	30.3	27.6	1.9	0.6	0.6	0.6	21.3	21.1	8.1	6.2
Llama3-8B-Instruct	52.0	30.4	26.5	25.6	1.6	0.5	0.5	0.5	16.9	18.7	7.3	6.1
Qwen2-7B-Instruct	46.7	27.6	23.3	28.2	1.6	0.5	0.4	0.6	13.9	17.1	5.1	5.6
Llama3.1-8B-Instruct	46.0	18.3	33.0	22.4	1.6	0.4	0.7	0.5	20.0	11.0	14.5	9.0
Phi-3-Mini-4K-Instruct	39.5	21.0	14.9	24.1	1.2	0.4	0.3	0.6	10.0	12.1	3.9	4.9
GPT-J-6B	35.1	9.2	5.9	30.1	0.9	0.1	0.1	0.7	5.0	5.9	1.8	4.5
Gemma-7B-Instruct	26.8	2.1	8.8	21.5	0.6	0.1	0.2	0.4	17.6	2.0	5.1	16.5

Table 4: Benchmark results sorted by overall occurrence rate. Higher metrics mean higher `SemSI` risk. Commercial closed-source models are put above open-source models. Experiments of GPT-o1 series are done at the end of September 2024 while other experiments are done at August 2024.

Conclusion

- We find that induced by simple natural questions, LLMs can output sensitive information at semantic level.
- We propose the definition and main categories of semantic sensitive information (SemSI).
- We build a dataset, SemSI-Set with 10,830 prompts and 9 SemSI labels, and a benchmark, SemSI-Bench with 3 types of metrics to systematically evaluate SemSI risk.
- We evaluate 25 LLMs and reveal several findings of the characteristics of SemSI.