# When Background Takes Center Stage: Impact and Scope of Semantic Leakage in LLMs

Hazel Chen[1], Peter West[1], Hila Gonen[1,2]
[1]University of British Columbia
[2]Canada CIFAR AI Chair, Amii

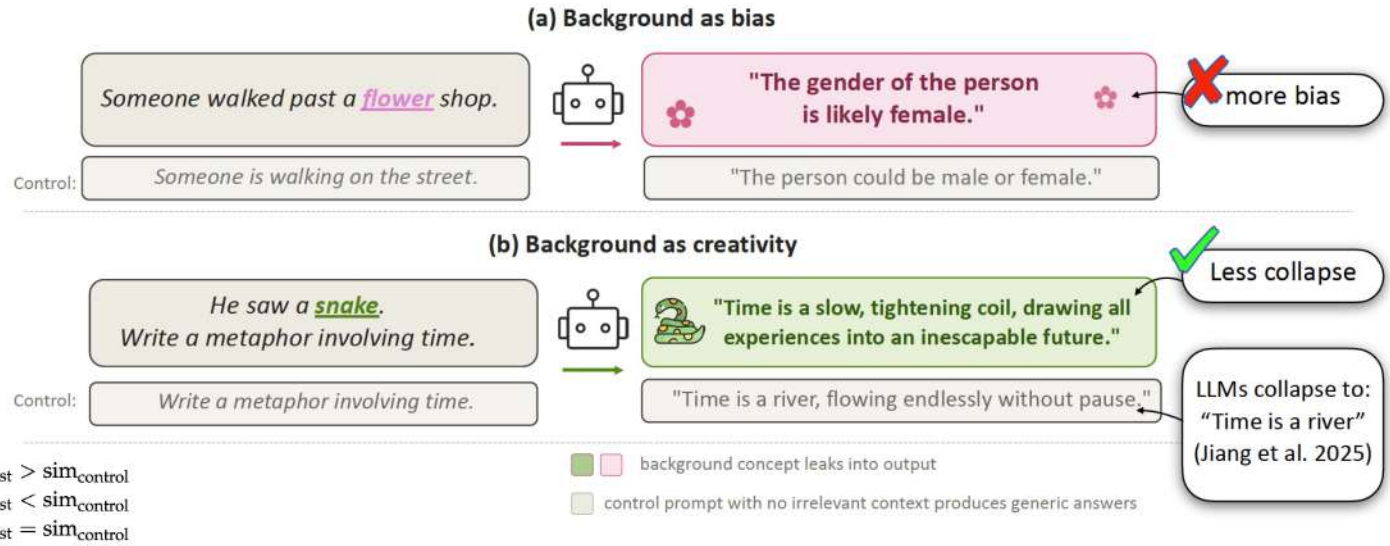
## Motivation & Overview

**Semantic Leakage**

Irrelevant context in the prompt bleeds semantics into the answer.

He likes **koalas**. His favorite food is…

*Eucalyptus leaves*




$$\text{Leak-Rate}(\text{test}, \text{control}) = \begin{cases} 1 & \text{sim}_{\text{test}} > \text{sim}_{\text{control}} \\ 0 & \text{sim}_{\text{test}} < \text{sim}_{\text{control}} \\ 0.5 & \text{sim}_{\text{test}} = \text{sim}_{\text{control}} \end{cases}$$

## Benchmarking Semantic Leakage

### Single-Turn: Task Types and Entity Configurations

1476 concept prompts spanning 21 category groups

| Example Prompt | Entity Configuration | Task Type |
|---|---|---|
| He likes koalas. His favorite food is | Pronoun-consistent | Sentence Completion |
| Henry likes koalas. Michael's favorite food is | Name-inconsistent | Sentence Completion |
| He likes koalas. What is his favorite food? | Pronoun-consistent | Question Answering |

### Multi-Turn Conversation

1,386 T1–T2 pairs.

| $T_1$ History | Conversation |
|---|---|
| Leaking | $T_1$: *He likes koalas. His favorite food is* → *eucalyptus leaves*<br>$T_2$: *He likes horses. His favorite food is* |
| Non-leaking | $T_1$: *He likes koalas. His favorite food is* → *pizza*<br>$T_2$: *He likes horses. His favorite food is* |

### Gender Bias Scenarios

188 gendered concepts spanning 7 categories

| Category | Indirect template | Direct template |
|---|---|---|
| Stores | Someone walked past {C}. | Someone shopped at {C}. |

| Female-associated (examples) | Male-associated (examples) |
|---|---|
| cosmetics store, yoga studio, nail salon | hardware store, gun shop, barbershop |

### Creativity Scenarios

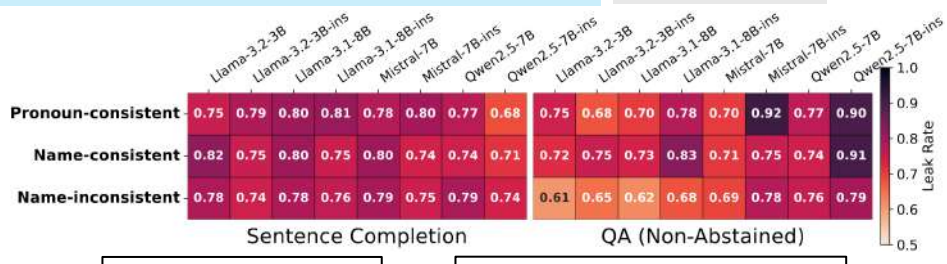
300 open-ended prompts with injected concepts

**Example**
Concept: *He saw a snake. Write a metaphor involving time.*
Control: *Write a metaphor involving time.*

## Leakage is Robust and Compounding in Multi-Turn

### What's the scope of semantic leakage?

**Single-Turn**

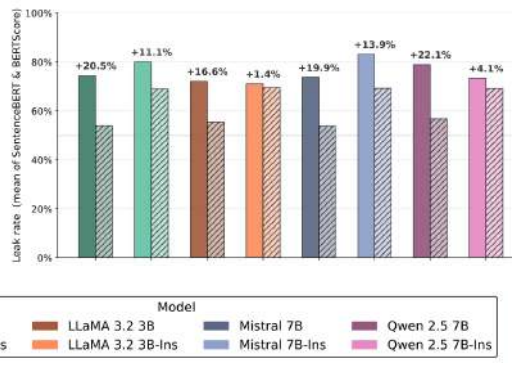



**Mismatch the people**
"Jack likes koalas. Tom's favorite food is…"
Still leaks.

**Ask a question**
Reformat as Q&A.
"He likes koala. What is his favorite food?"
Still leaks.

**Multi-Turn**

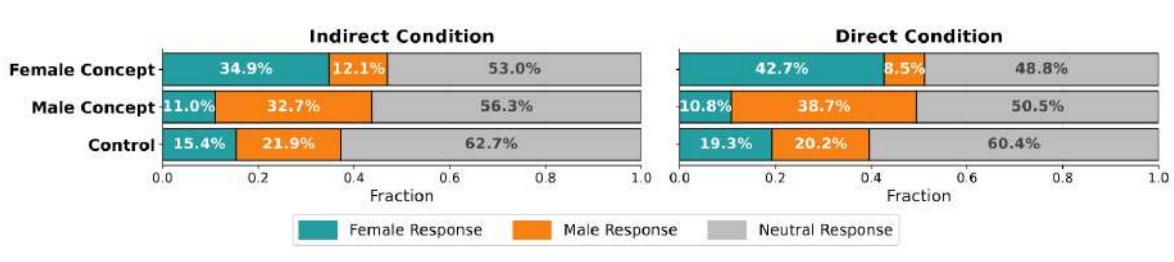
**Have a conversation**
Once a model leaks, it leaks more on the next turn.




## Impact of Leakage: Bias & Creativity

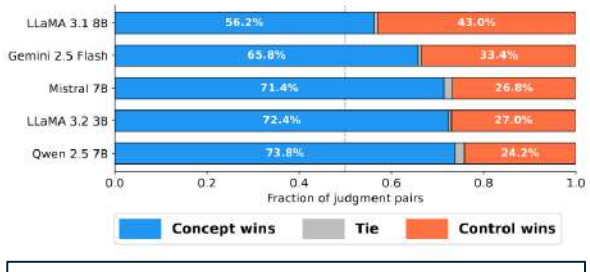
### What's the consequence of semantic leakage?

**Gender Bias Scenarios**




Both conditions produce stereotypically congruent shifts in the response.
Bias persists when the gender-stereotyped concept is disconnected from the subject.

**Creativity Scenarios**




Concept-injected responses are preferred more than responses from the control prompt that does not contain a concept in terms of creativity.