



Disentangling In-Weight and In-Context Learning for Contributive Attribution

Quoc-Huy Trinh, Lin Zhu, Sebastian Szyller

Motivation

Prompt:
Which art movement was Claude Monet a founder of ?

Context	Score	GT
Leonardo da Vinci painted the Mona Lisa during the Renaissance.	0.1	irrelevant
Later, Claude Monet co-founded the Impressionism movement to capture light and fleeting moments.	0.3	supporting
Impressionist painters typically worked outdoors to capture real-time atmospheric changes.	0.5	related

Response: Claude Monet is the founder of Impressionism.

Not in weight (blue) In weight (yellow)

1. Attribution methods **fail to distinguish** between in-weight knowledge (IW) and context (ICL)
2. Context-level scores become **hard to interpret** when the context overlaps with training data

How do context attribution methods handle what the LLM already knows

Benchmarking attribution

Three new **metrics** and new dataset

- **BCS** Does model rely on top-k context units?
- **CAC** Do attribution rankings stay stable after finetuning?
- **SSP-ICL** Segments predicted as ICL that are truly ICL?
- **SSP-IW** Segments predicted as IW that are truly IW?

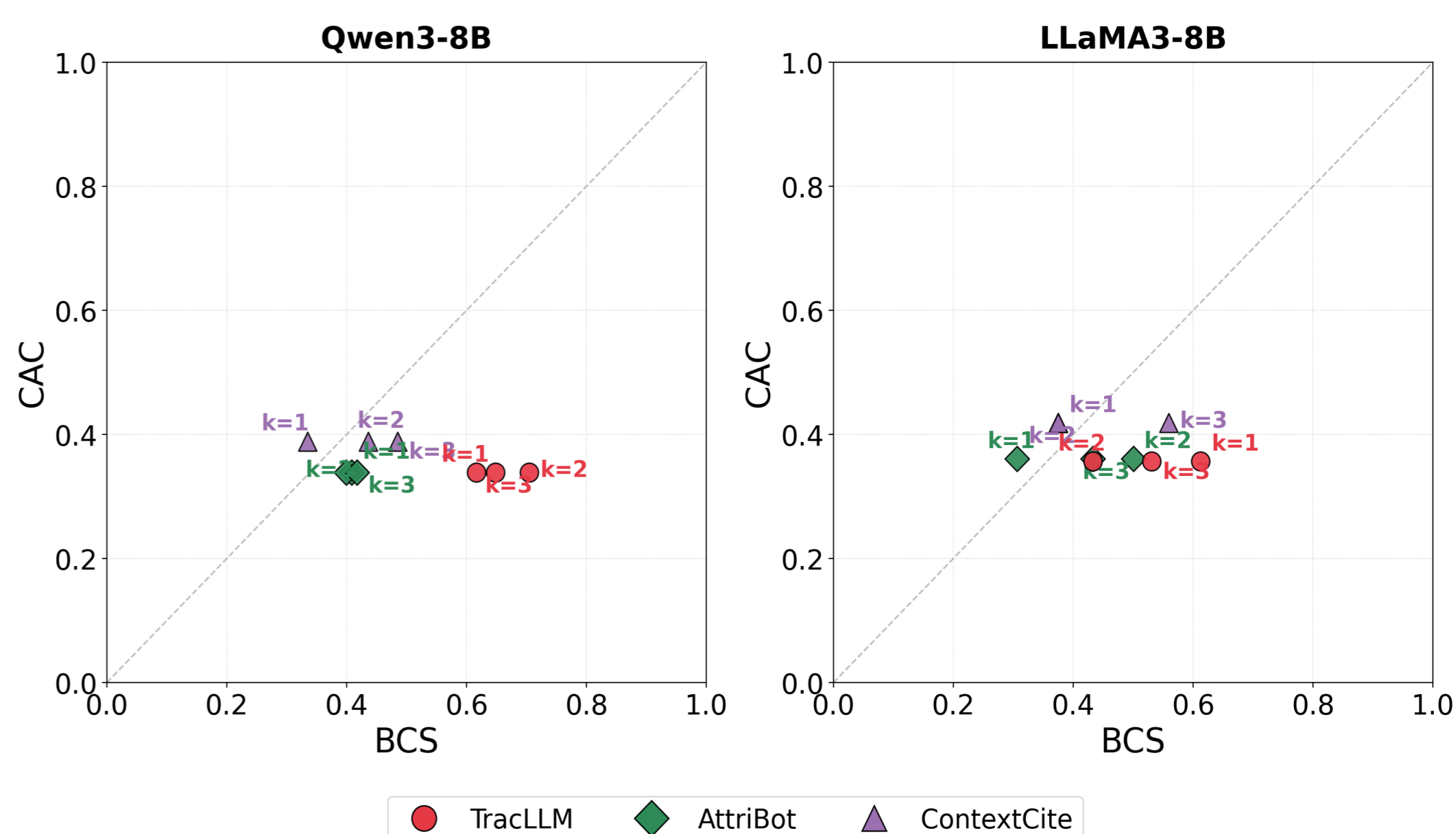
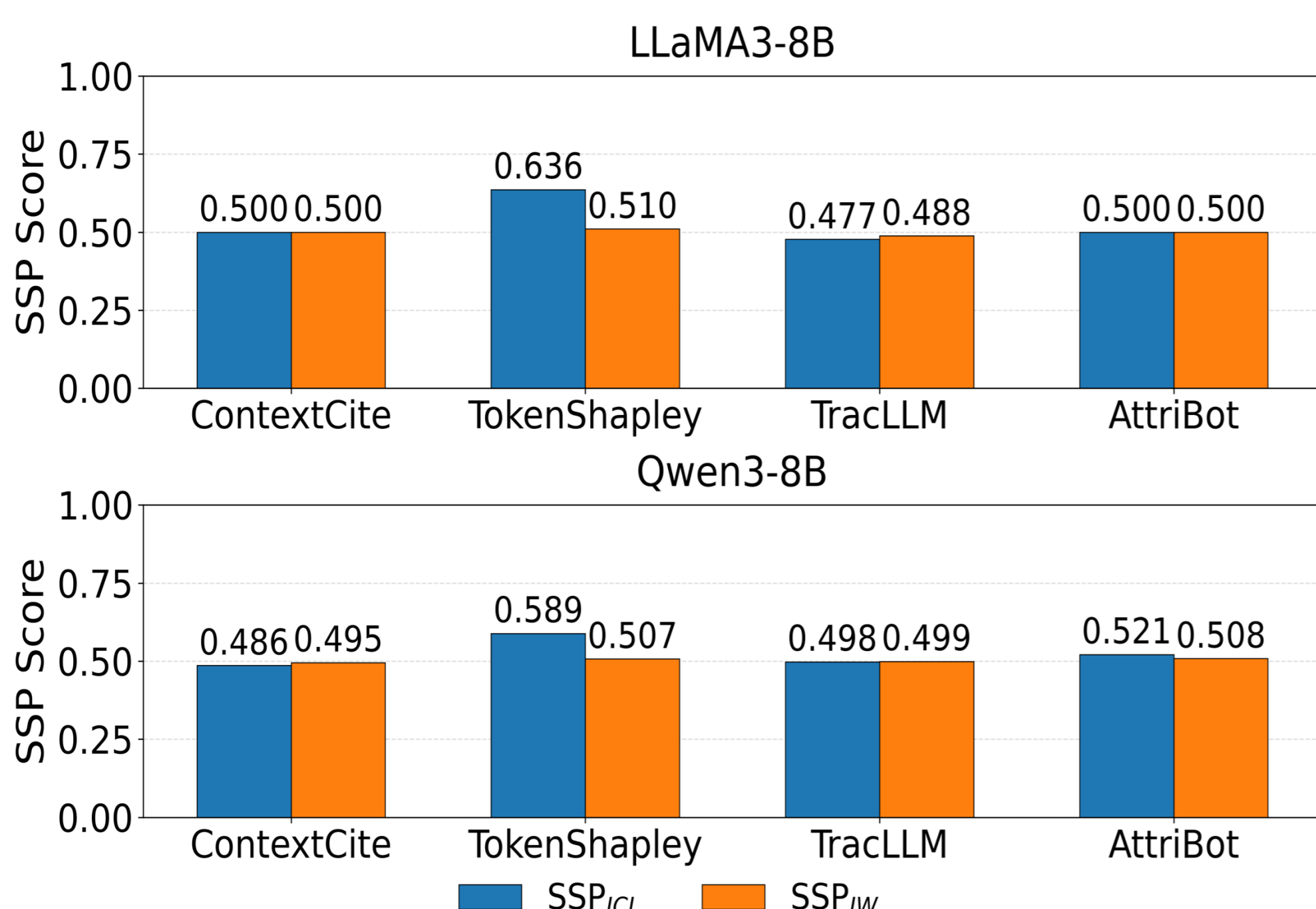
WMDP-Cyber++

- 1,987 cybersecurity multiple-choice questions
- each paired with **mixed context**
- **IW** segments used for finetuning
- held out **ICL** segments used during inference
- ground-truth provenance labels by **construction**

Result

SSP scores stay **near random**

IW knowledge shifts attribution (CAC); ICL attribution remains limited (BCS)



Open questions: how can new attribution methods distinguish between IW and ICL?

Work in progress