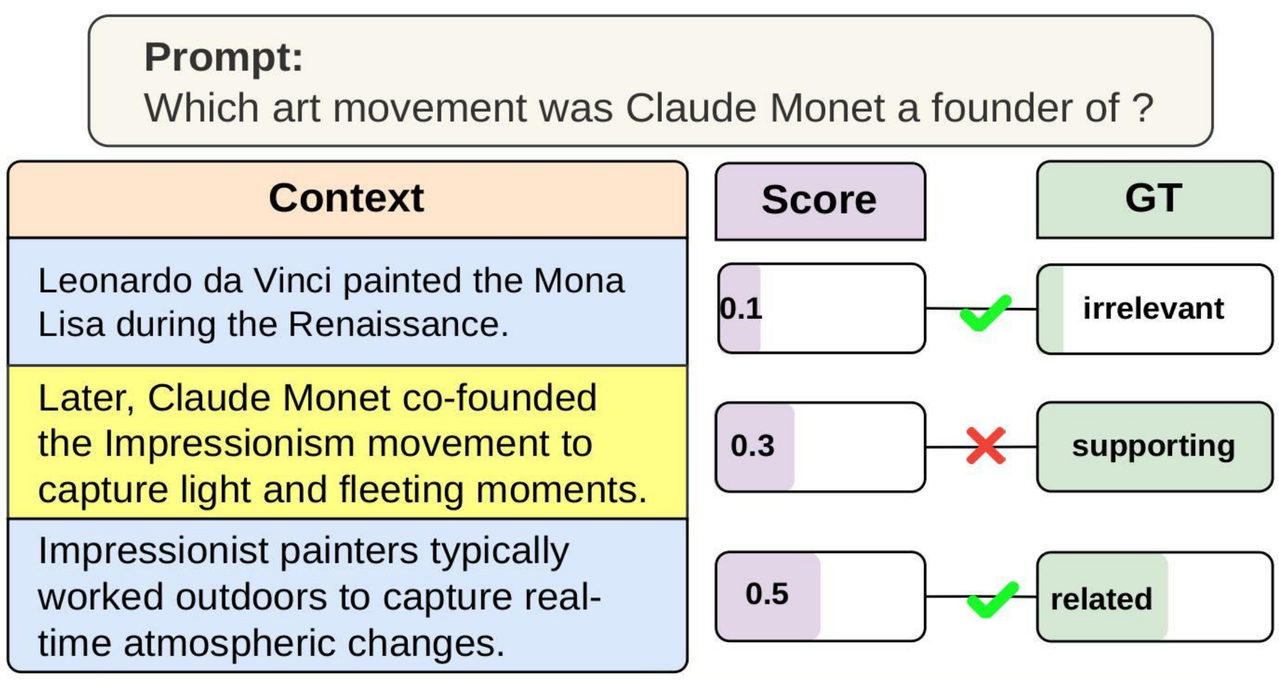
Aalto University
School of Science

taclab.aalto.fi

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

Quoc-Huy Trinh, Lin Zhu, Sebastian Szyller

Motivation



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

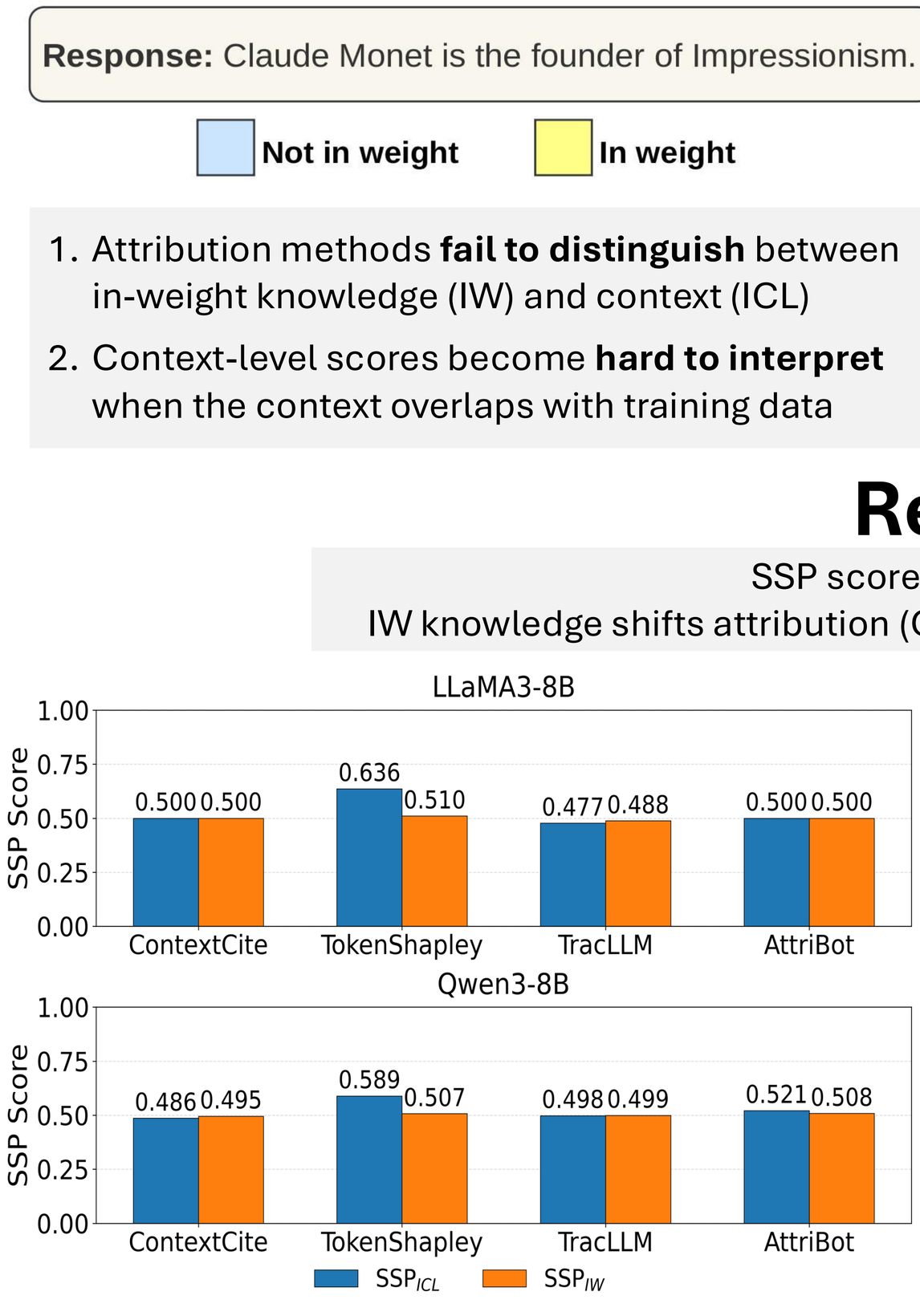
- **BCS** Does model rely on top-k context units?
- **CAC** Do attribution rankings stay stable after finetuning?
- **SSP-ICL** Segments predicted as ICL that are truly ICL?
- **SSP-IW** Segments predicted as IW that are truly IW?

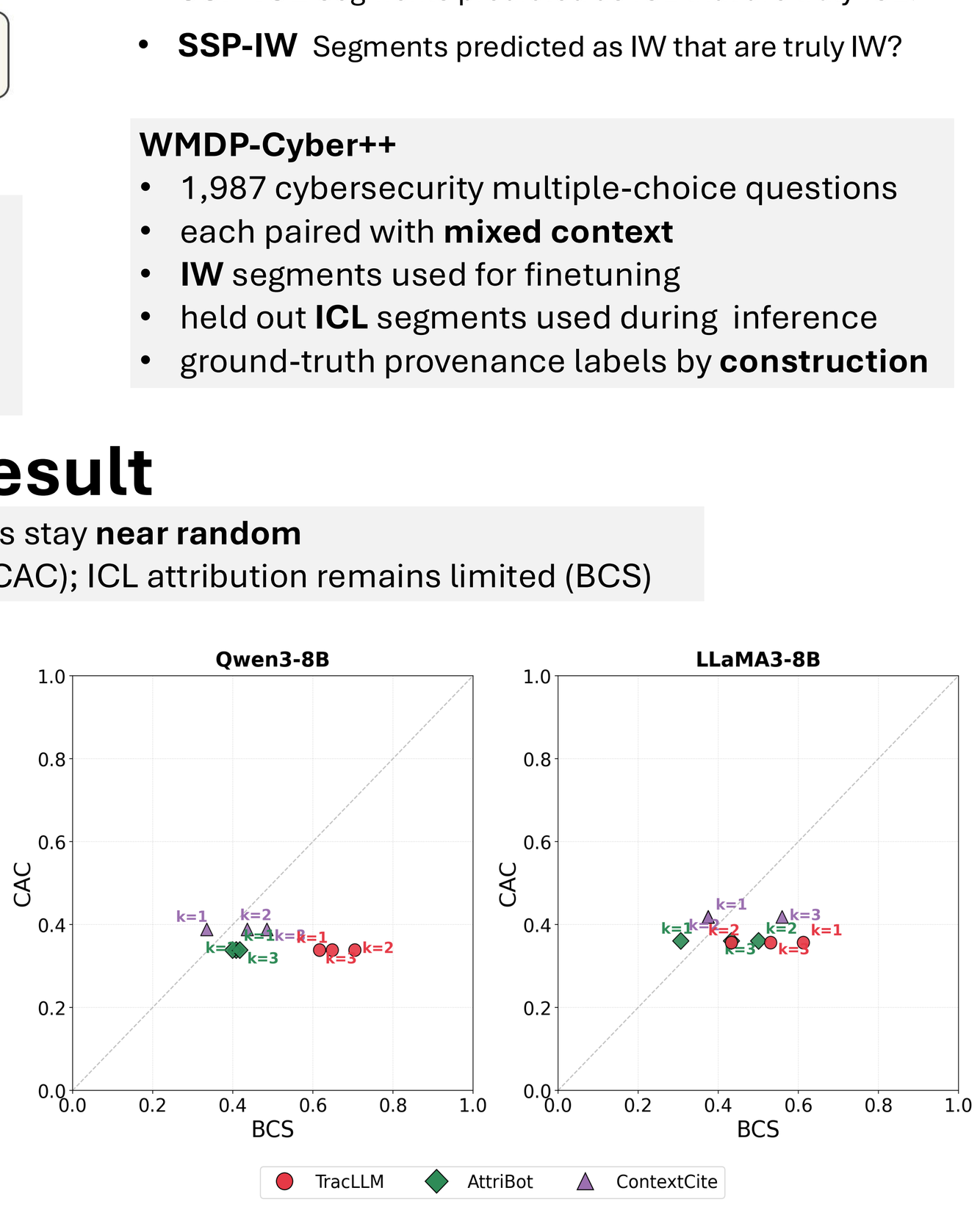
WMDP-Cyber++

- 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