

CardiO: Predicting Cardinality from Online Sources

Shrestha Ghosh, Simon Razniewski*, Damien Graux†, Gerhard Weikum

Max Planck Institute for Informatics, Germany *Bosch Centre for AI, Germany †Trinity College Dublin, Ireland

Motivation

Counts from top SE snippets returned for count questions

how many AI startups in Europe?
(well-defined, but not very popular)

- Few relevant counts
Europe has 1,157 AI startups
- Counts from popular subsets
The UK .. with 500 AI startups Europe has 150 generative AI startups
- Counts from popular peers
The US has 1,751 AI startups

how many programmers in Europe?
(vaguely-defined, quite popular)

- Frequent relevant estimates
estimated 6.1 million developers approx. 6.1 million software developers
- Noise
500.000 developers short in job market
- Time-variance
July 2023 approx. 5.7 million software developers in Europe

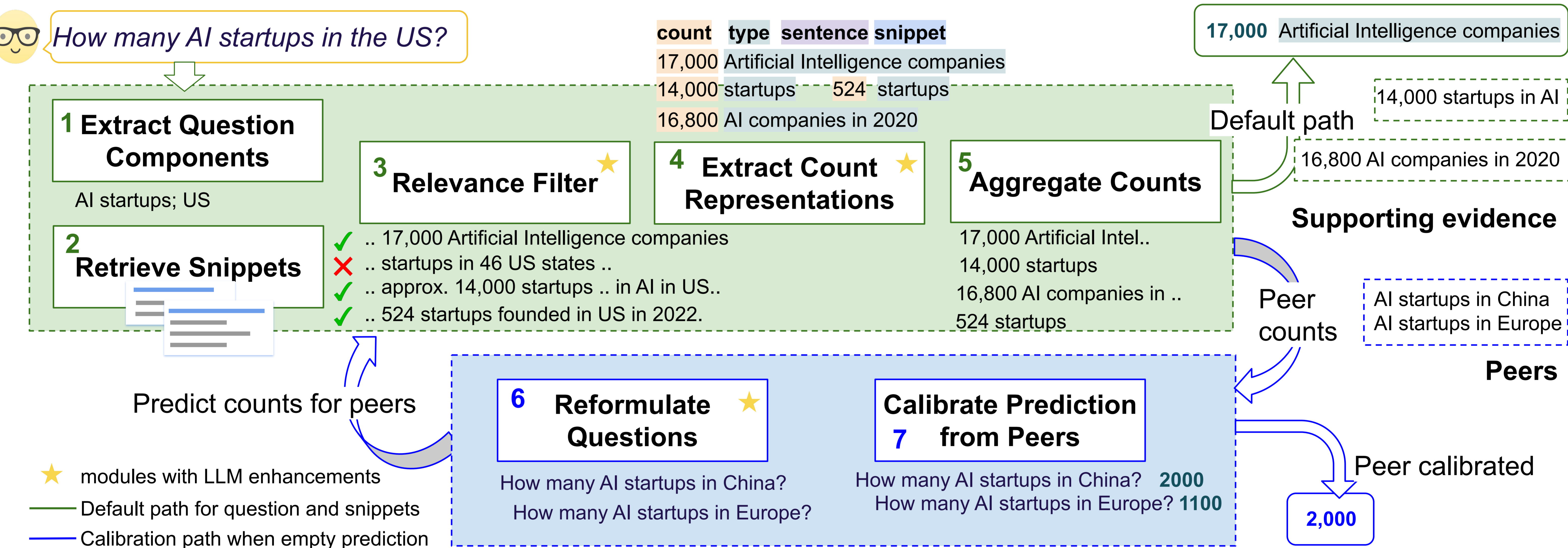
More count questions

how many beaches are Blue Flag certified?
how many films produced by Warner Bros?
how many lakes are there in the world?

- **Incompleteness and bias in online world**
Different sources focus on different angles.
Topic popularity and crispness affect count availability.
- **Distribution of counts on the Web**
Many relevant contexts with close counts; noise.
- **Explanatory evidence over direct answers**
Context, instances, and snippets increase user comprehension.
- **LLMs have high precision but lack traceability**
LLM-only approach has no traceability and relies on (outdated) parametric knowledge.
Retrieval-augmentation does not guarantee provenance.

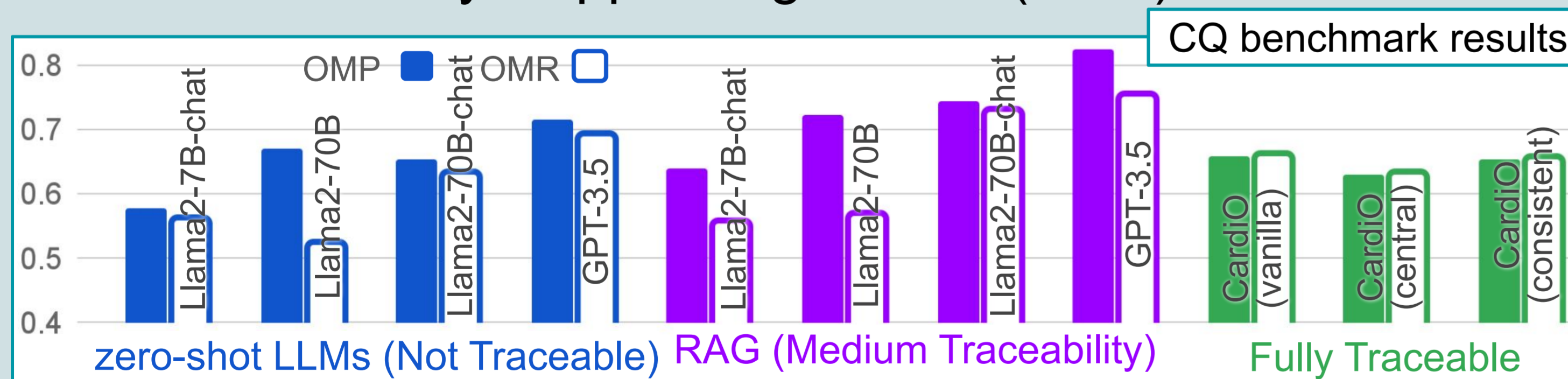
- CardiO is a lightweight and modular framework.
- Provides traceability with tangible evidence.
- Makes distribution-aware prediction by using supporting evidence

CardiO Framework

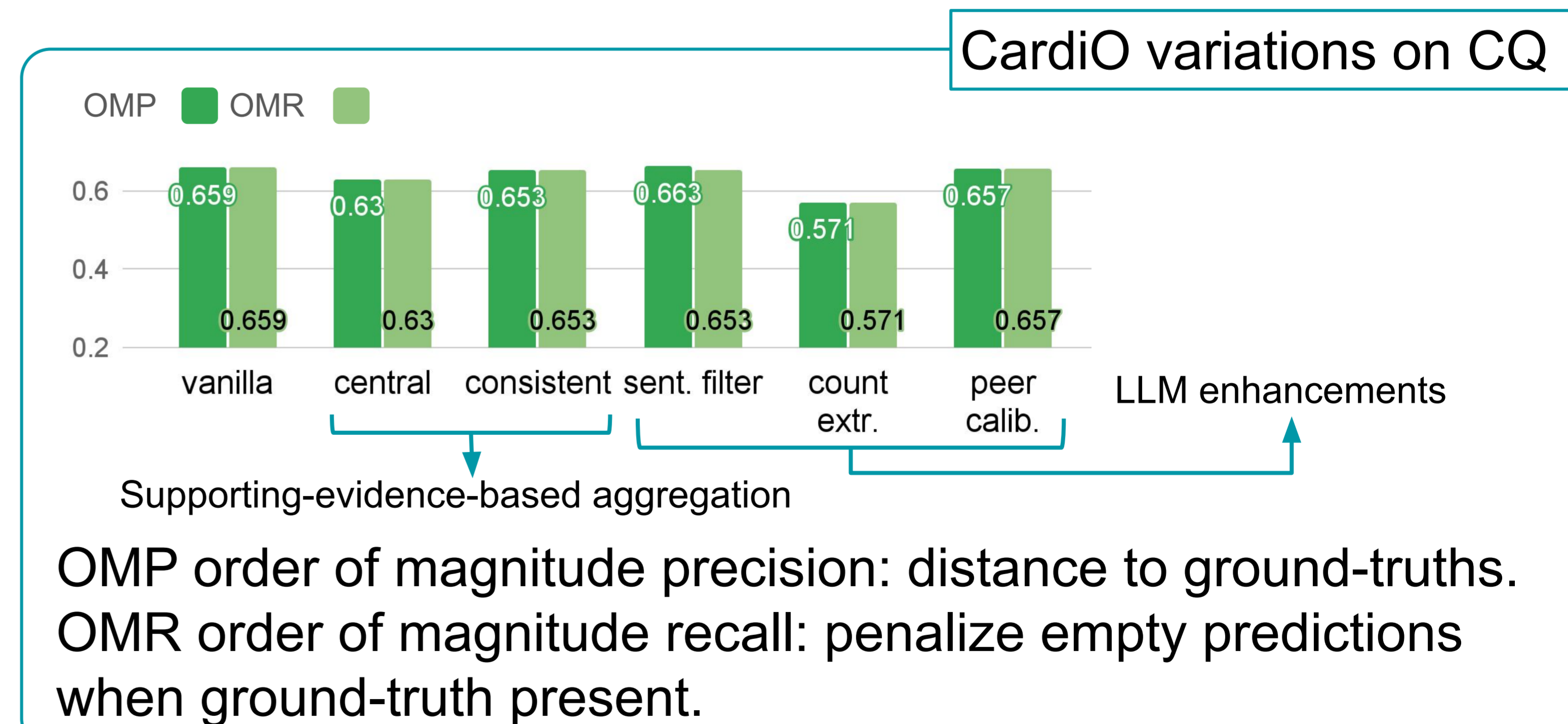


Results

- Tradeoff between untraceable zero-shot LLMs and resource-heavy snippet-augmented (RAG) LLMs.



- Supporting-evidence-based aggregation strategies.
- Controlled LLM enhancements in three modules:
 - Sentence filter based on relevance
 - Count representation extraction
 - Peer calibration from entities with close cardinalities



- **Cardinality Benchmarks**
- CQ (ours): 500 questions; annotated question and ground-truth properties.
- NQ: 84 count questions from Natural Questions dataset.
- CoQuAD: 312 count questions.



Dataset & Results

[1] Ghosh et al. 2023. Class Cardinality Comparison as a Fermi Problem. In WWW
 [2] Ghosh et al. 2022. Answering Count Queries with Explanatory Evidence. In SIGIR
 Our research page on Count Knowledge: <https://www.mpi-inf.mpg.de/count-knowledge>
 Dataset and Results: <https://github.com/ghoshs/CardiO>