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

Counts from top SE snippets returned for count questions

- how many AI startups in Europe? (well-defined, but not very popular)
- how many AI startups in the US?
- how many programmers in Europe? (vaguely-defined, quite popular)
- how many AI startups in China?
- how many beaches are Blue Flag certified?
- how many AI startups in Europe? (vaguely-defined, quite popular)
- how many films produced by Warner Bros?
- how many lakes are there in the world?

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

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LLM-only approach has no traceability and relies on (outdated) parametric knowledge.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 Framework

How many AI startups in the US?

1. Extract Question Components
   AI startups; US

2. Retrieve Snippets
   modules with LLM enhancements
   Default path for question and snippets
   Calibration path when empty prediction

3. Relevance Filter
   .. 17,000 Artificial Intelligence companies
   .. startups in 46 US states ..
   .. approx. 14,000 startups .. in AI in US..
   .. 524 startups founded in US in 2022.

4. Extract Count Representations
   Count type sentence snippet
   17,000 Artificial Intelligence companies
   14,000 startups ... 524 startups
   16,800 AI companies in 2020
   Peer calibrated

5. Aggregate Counts
   17,000 Artificial Intelligence companies
   14,000 startups
   16,800 AI companies in ...
   524 startups

6. Reformulate Questions
   How many AI startups in China?
   How many AI startups in Europe?

7. Calibrate Prediction from Peers
   How many AI startups in China? 2000
   How many AI startups in Europe? 1100

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

CardiO variations on CQ

Cardinality Benchmarks

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

Our research page on Count Knowledge: https://www.mpi-inf.mpg.de/count-knowledge
Dataset and Results: https://github.com/ghoshs/CardiO