How well can LLMs “chat” with Knowledge Graphs?

Benchmarking the Abilities of Large Language Models for RDF Knowledge Engineering Tasks

Johannes Frey, Lars-Peter Meyer,

Felix Brei, Natanael Arndt, Kurt Junghanns, Kirill Bulert, Sabine Gründer-Fahrer, Michael Martin

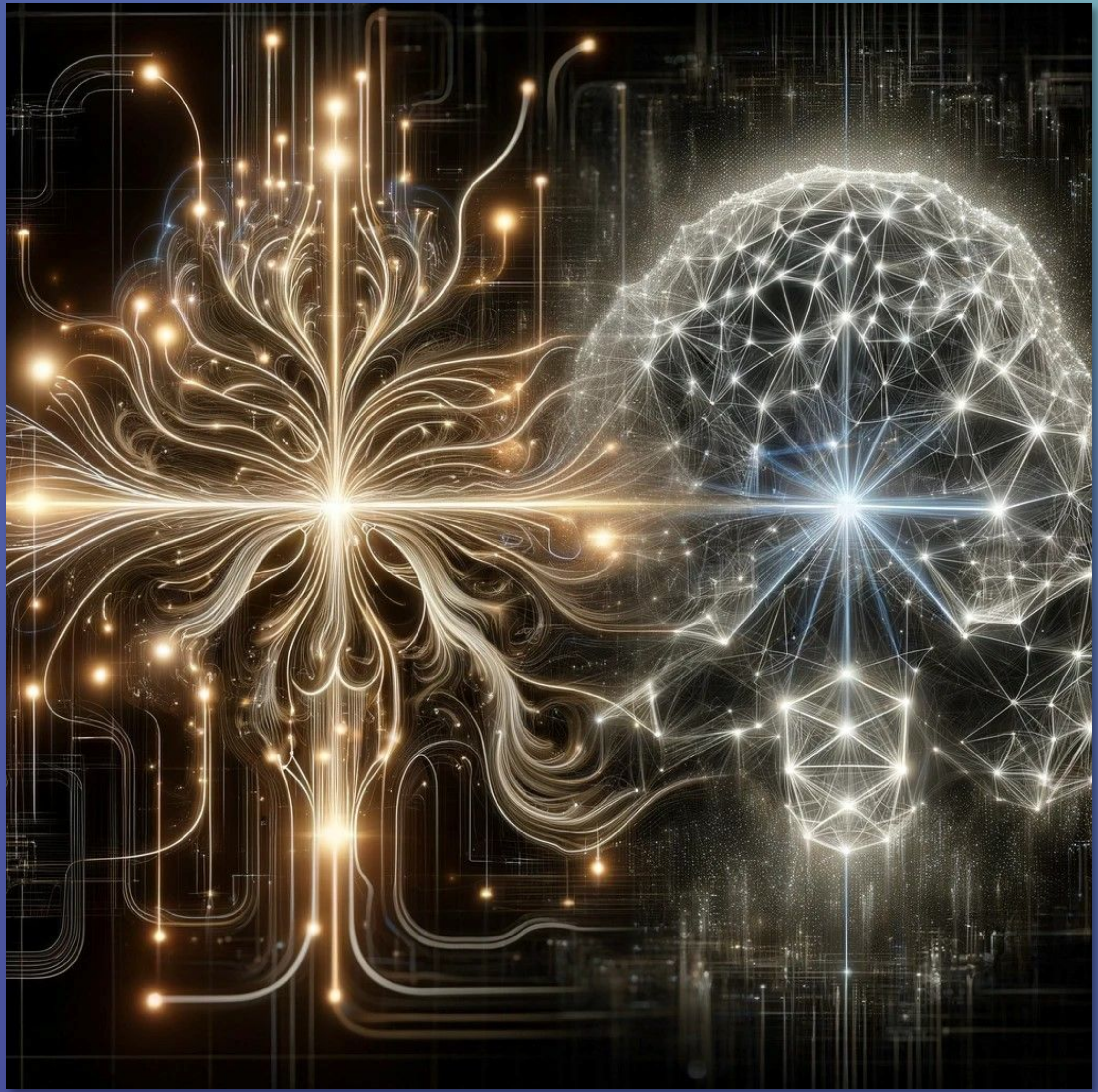
Institute for Applied Informatics & Leipzig University

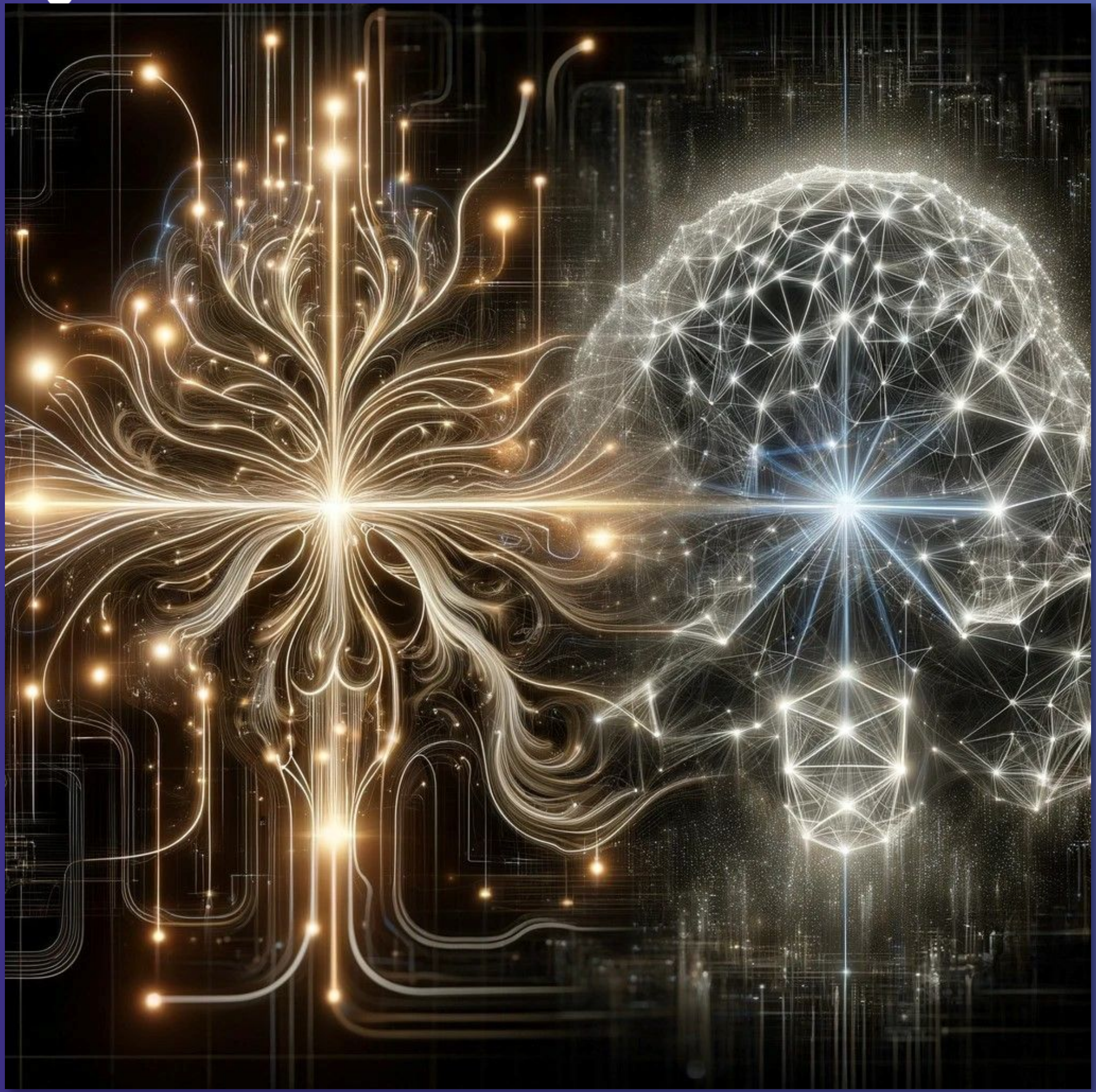
Supported by grants
FKZ: 13XP5116B FKZ: 01MK21007A
FKZ: 13XP5119F FKZ: 1MK22001A

LLMs □ KGs

Employ data from KG to improve answers:

- Reduce hallucinations
- Cover more recent data
- Cover long tail or private data not covered in model
- Look up of precise data (e.g. DOIs of papers)





LLMs □ KGs

Assist in Knowl. Eng. Tasks:

- Knowledge Extraction
- Mapping Generation
- Ontology / Entity matching
- Error curation
- …

Motivation

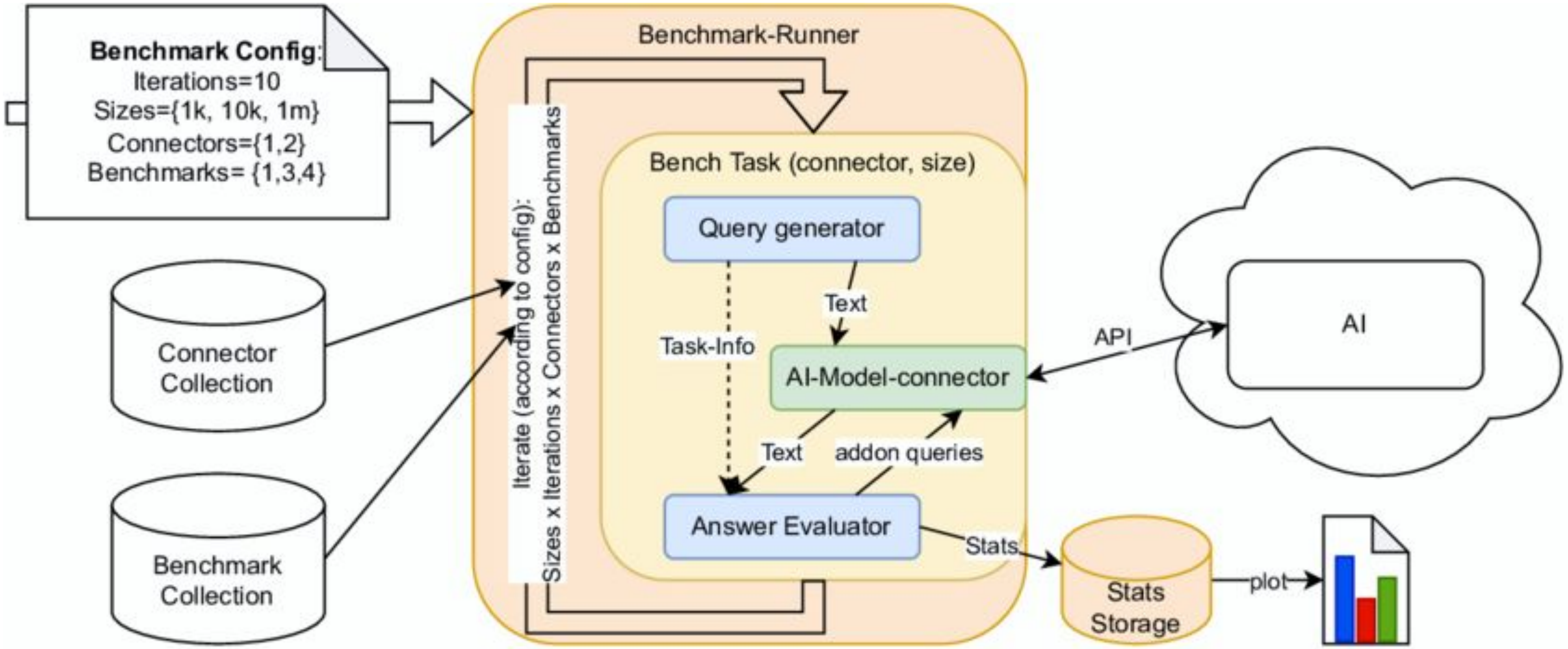
□need I/O Interfaces between LLMs and KGs

□ Intuitive choice: Turtle and SPARQL

- Natural way of encoding facts and relationships between things (“SPO style”)
- Well **standardized**
- Lots of consistent training data accessible on the web

Method: LLM-KG bench Framework

- **Automated LLM assessment framework for KGE related capability testing**
  - Connectors for commercial models Claude, OpenAI, Gemini and GPT4all
- **Currently 6 task types: 5 Turtle and 1 SPARQL**
- **2 scalable tasks can be configured in problem size, SPARQL is instance based**



Benchmark Tasks: RDF skill levels

Task T1: Find a connection between Nodes in Turtle

Task T2: Find syntax errors in Turtle

Task T3: Generate Sample Person Graphs

Task T4: Identify most known Person

Task T5: Extract Data from 3D Printer Factsheet

Task T6: Text2SPARQL

	T1	T2	T3	T4	T5	T6
Turtle Read	+	++	/	+	/	/
Turtle Write	/	++	+	/	++	+
Graph Understanding	++	/	+	++	++	++
Vocabulary Knowledge	/	/	/	/	+	+

Benchmark Tasks: Setup

- *static* tasks run with 20 iterations
- *scalable* tasks run for 8 different problem sizes (20 iterations each)
- Task sizes configured via byte limit (1000, 2000,… 8000)
- Scalable tasks derive a custom problem size based on the limit trying to approximate it (by estimating the sum of prompt and response length in chars)

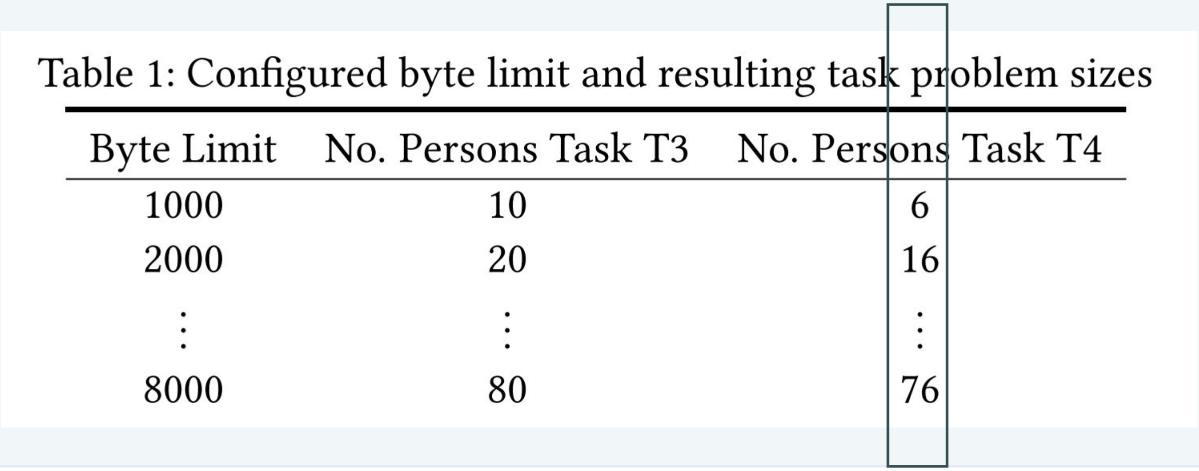
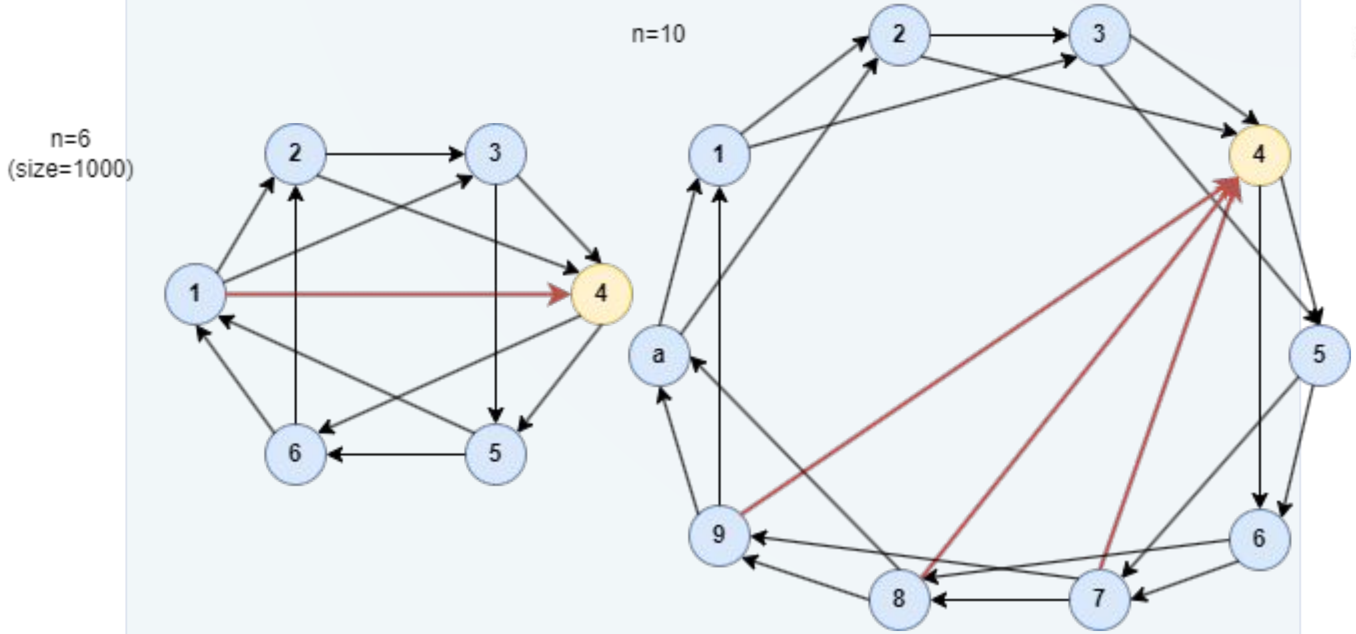


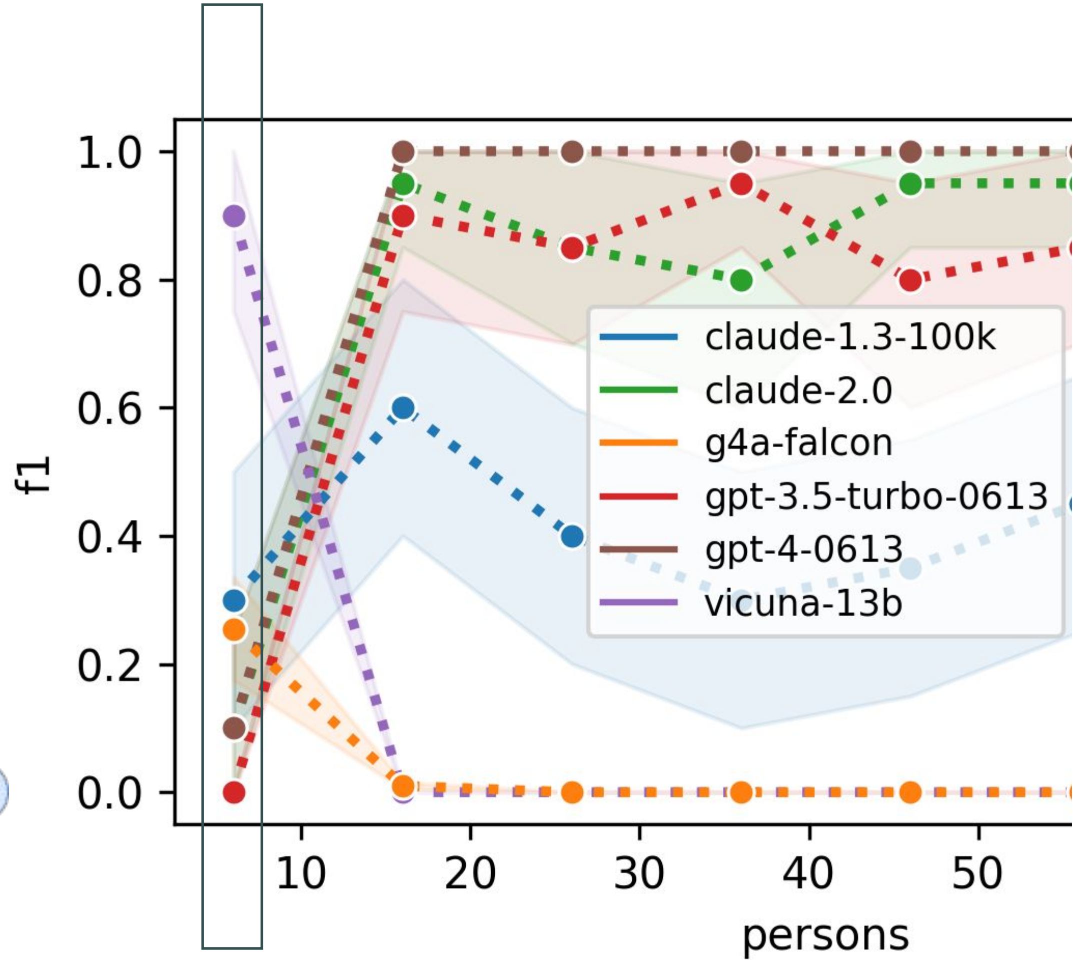
Table 1: Configured byte limit and resulting task problem sizes

Byte Limit	No. Persons Task T3	No. Persons Task T4
1000	10	6
2000	20	16
⋮	⋮	⋮
8000	80	76

T4: Most-known Person - Special Case – 6 Persons

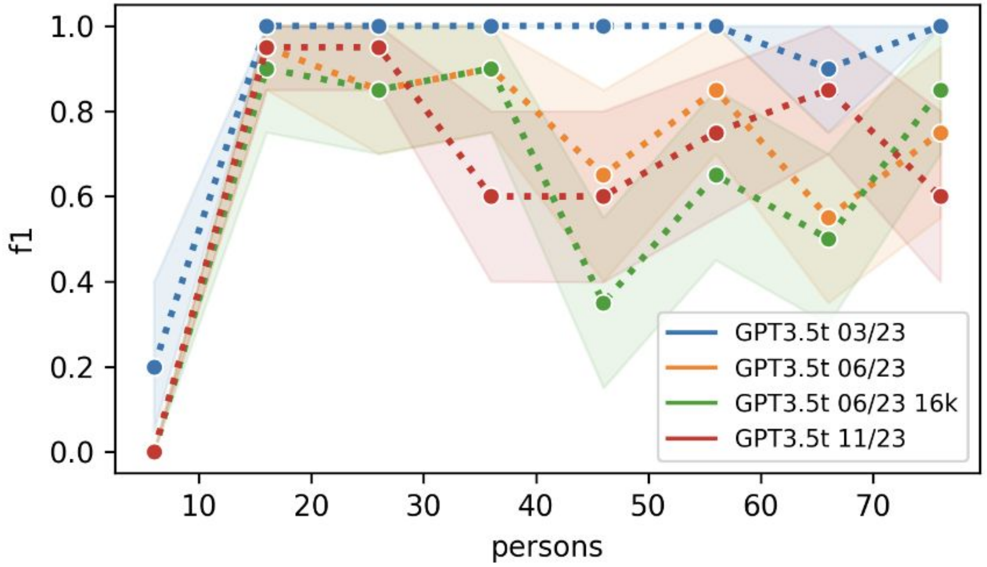
- Special case #outlinks of person1 are equal to #inlinks for person4
- All models besides Vicuna often confuse ingoing vs. outgoing links

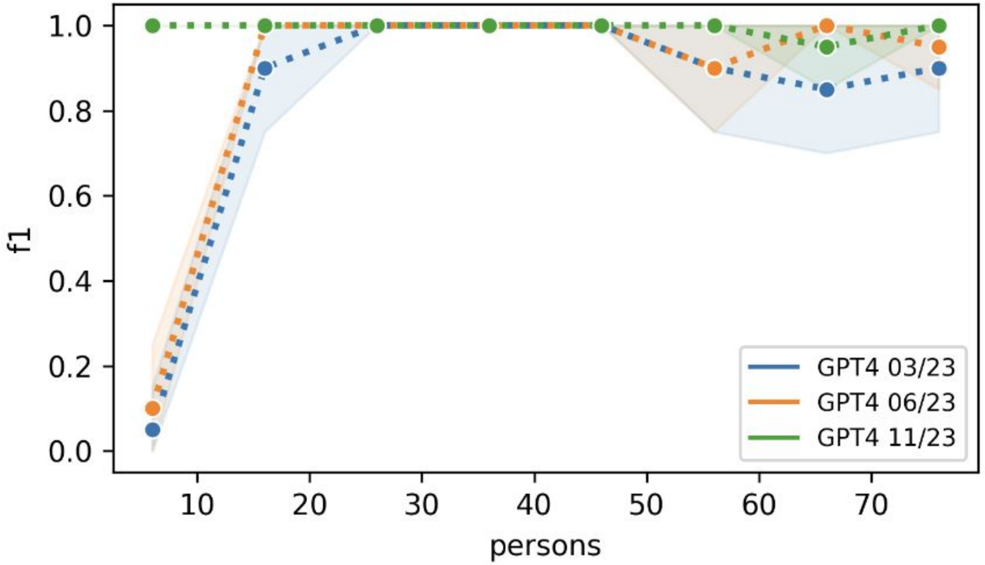




T4: GPT Evolution

- GPT4 11/23 first model that can handle the tricky case
- GPT3.5 03/23 outperforms all newer GPT3.5 models!

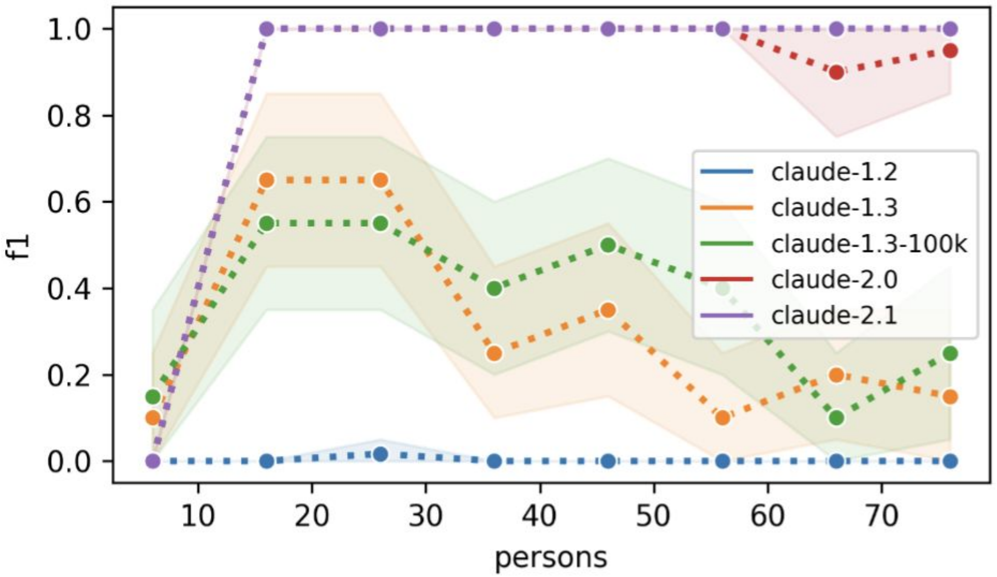


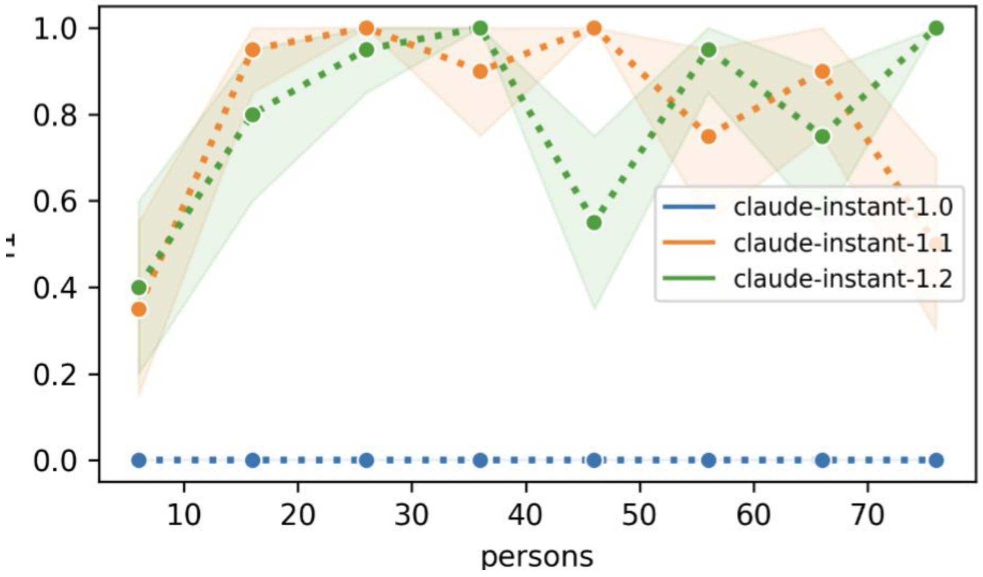


T4: Claude Evolution

- Instant models adhere better to given output requirements
- No model is reliable for tricky case

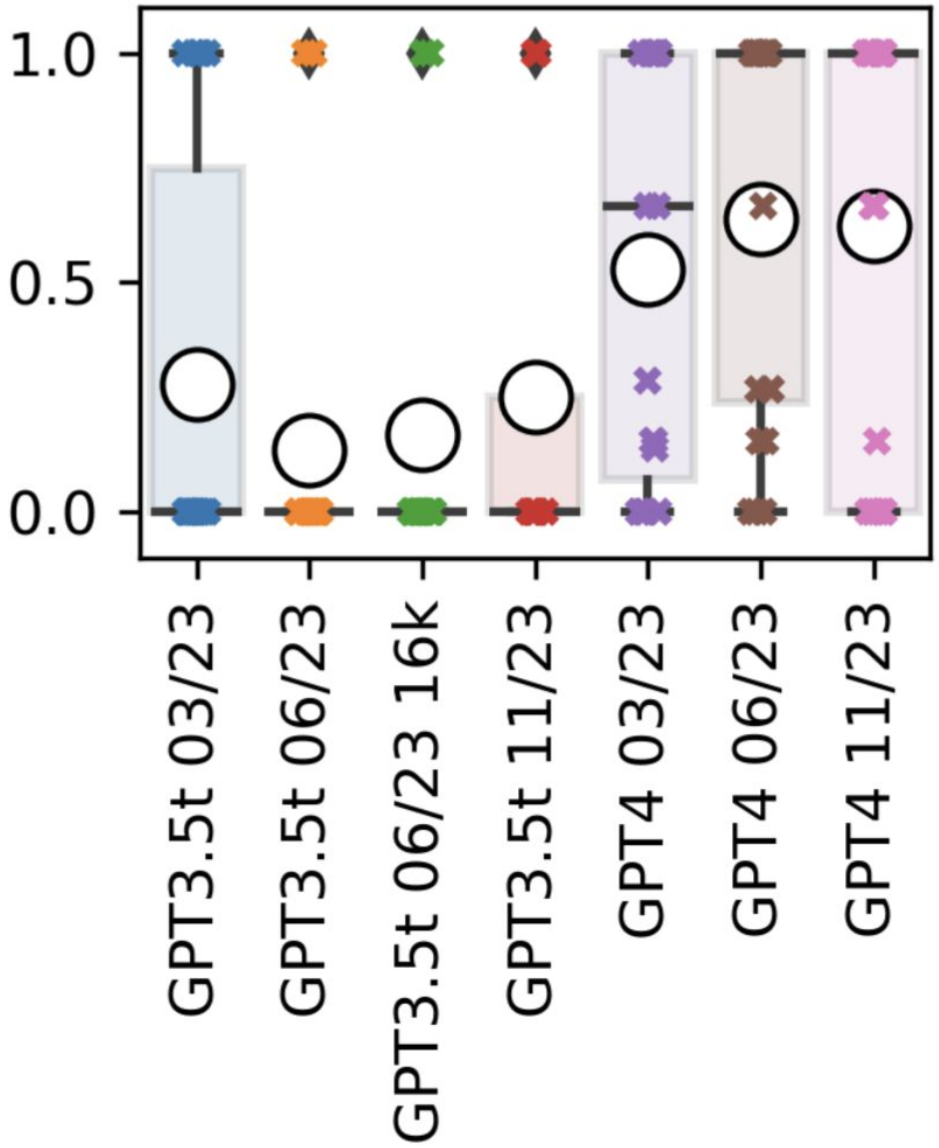
□ Claude 1.x /2.x missing understanding of edge direction





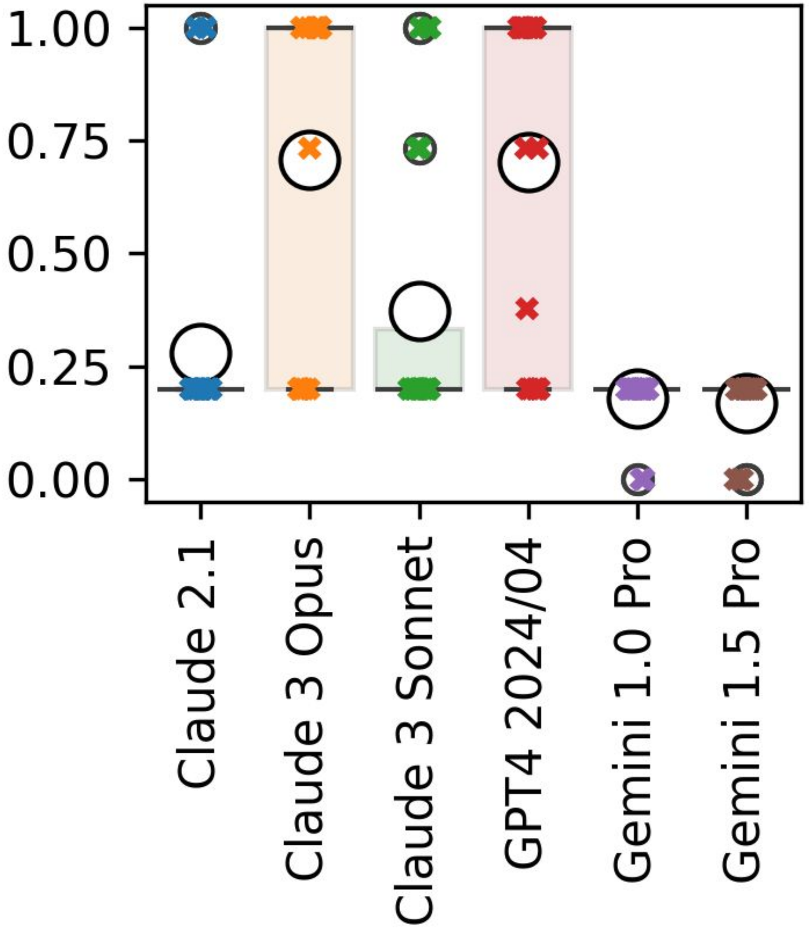
T6: Text2SPARQL

- Task: convert LC-Quad-question to SPARQL query
- Using multi-shot prompting with feedback loop on errors and reevaluate next answers
- Mapping between IRIs and Labels occuring in gold query are provided
- Claude versions had only one correct result



T6: „Nightly“ Results

- **Updated run with new models**
- **Combined score:**
  - =0: syntax not correct
  - >=0.2: syntax correct
  - =1.0: syntax and result correct

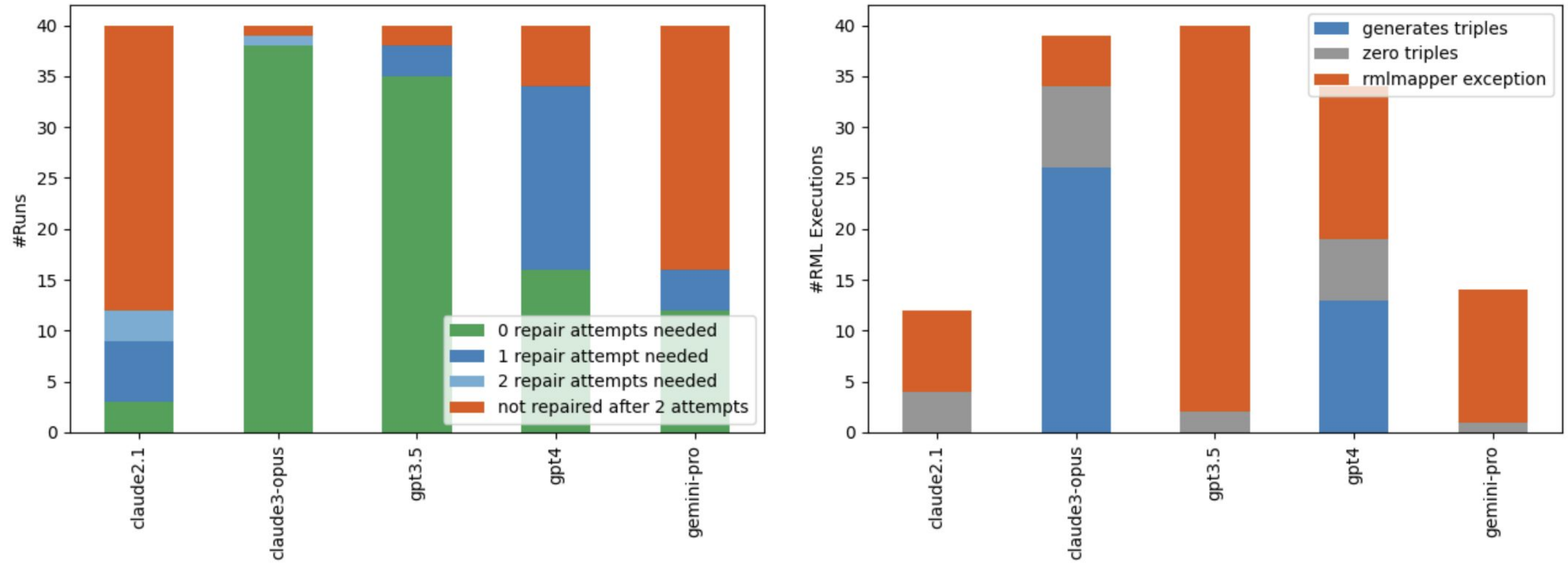


Selected Findings & Conclusions

- **Trend for newer/latest version of commercial models**
    - outperform their forerunners
    - but have a tendency to give extra text in the response or markdown ticks
        - useful for assisting humans, but problematic when interfacing with (RDF) tools

- **Selected GPT4all models not useful for KGE tasks at current stage**

Teaser: RML Mapping generation



(a) RML Turtle Syntax Validity

(b) Mapping Soundness

Discussion point

- How to prevent that LLMs learn Benchmark results for public benchmarks?

Thank you

CONTACT

Lars-Peter Meyer

lpmeyer@infai.org

Johannes Frey

frey@infai.org

KISS
KI-GESTÜTZTES RAPID SUPPLY NETWORK

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

References

5 Towards self-configuring Knowledge Graph Construction Pipelines using LLMs-A Case Study with RML

M Hofer, J Frey, E Rahm

Fifth International Workshop on Knowledge Graph Construction@ ESWC2024

4 Assessing the Evolution of LLM capabilities for Knowledge Graph Engineering in 2023

J Frey, LP Meyer, F Brei, S Gründer-Fahrer, M Martin,

ESWC2024

3 Benchmarking the Abilities of Large Language Models for RDF Knowledge Graph Creation and Comprehension: How Well Do LLMs Speak Turtle?

J Frey, LP Meyer, N Arndt, F Brei, K Bulert

Deep Learning for Knowledge Graphs @ ISWC2023 3559 (CEUR WS Proceedings …

2 Developing a Scalable Benchmark for Assessing Large Language Models in Knowledge Graph Engineering

LP Meyer, J Frey, K Junghanns, F Brei, K Bulert, S Gründer-Fahrer, ...

Semantics 2023 3526 (CEUR WS Proceedings SEMPDS 2023)

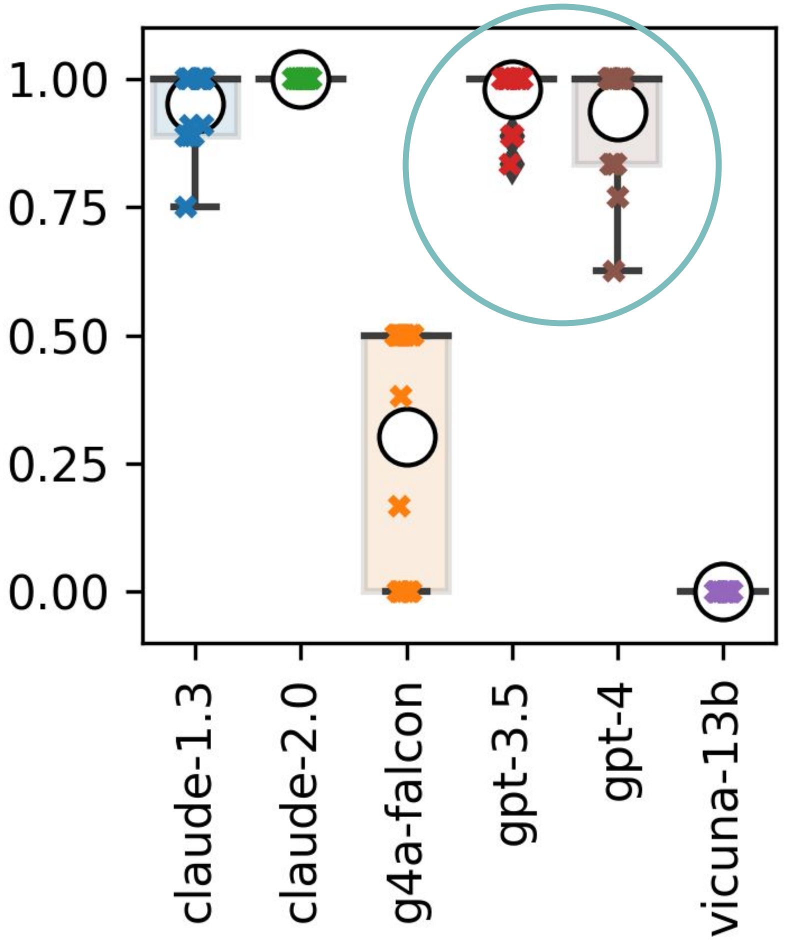
1 LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT

LP Meyer, C Stadler, J Frey, N Radtke, K Junghanns, R Meissner, ...

AI Tomorrow @ Data Week

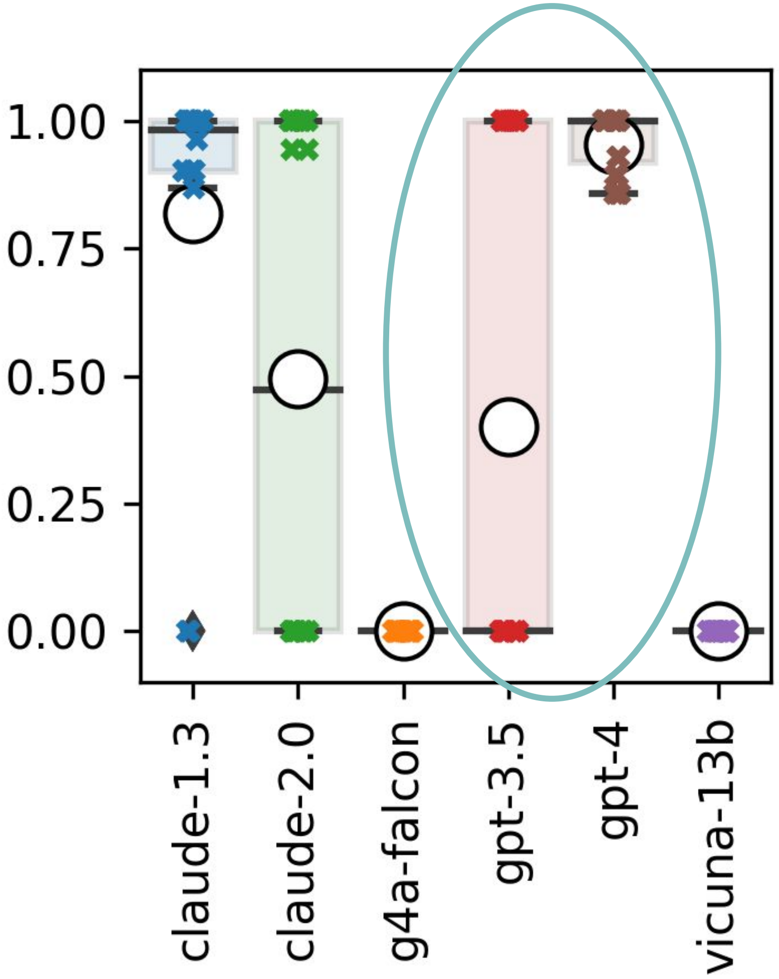
Task T1: Evaluation findings & F1 score

- Claude 2 perfect
- GPT4 sometimes lists properties
- GPT3.5 and Claude 1.3 miss occ. a resource
- Falcon has basic understanding, but sometimes lists only Anne & Bob
- Vicuna mostly reports "This is the shortest connection from anne to bob"



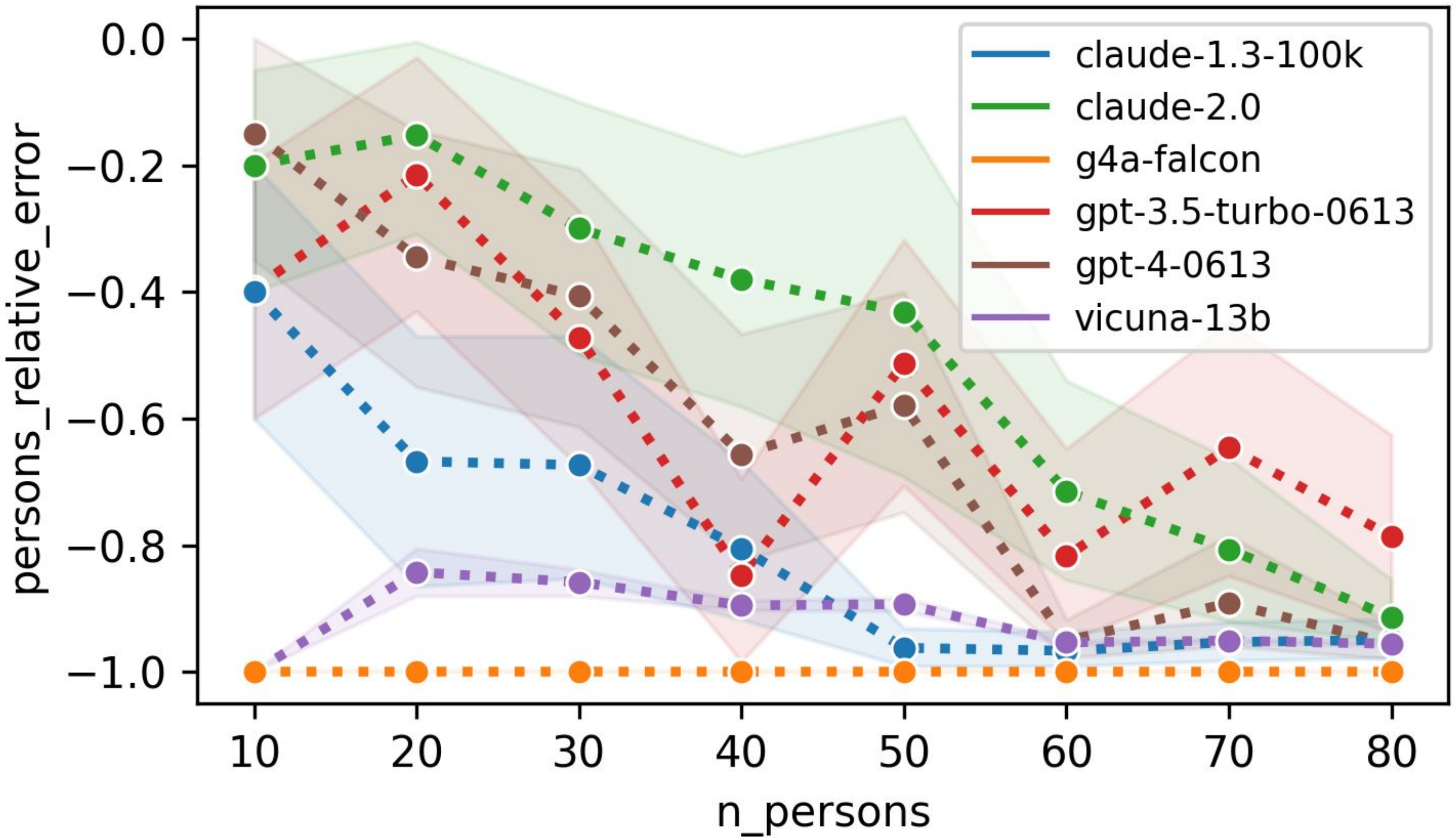
Task T2: Find syntax errors in Turtle - F1 score

- GPT4 is best followed by Claude 1.3
- GPT3.5 “all or nothing” - often claims the file to be correct and returns no Turtle
- Claude 2 fails in returning plain turtle
- Falcon explains content or claims no error
- Vicuna replies with empty string



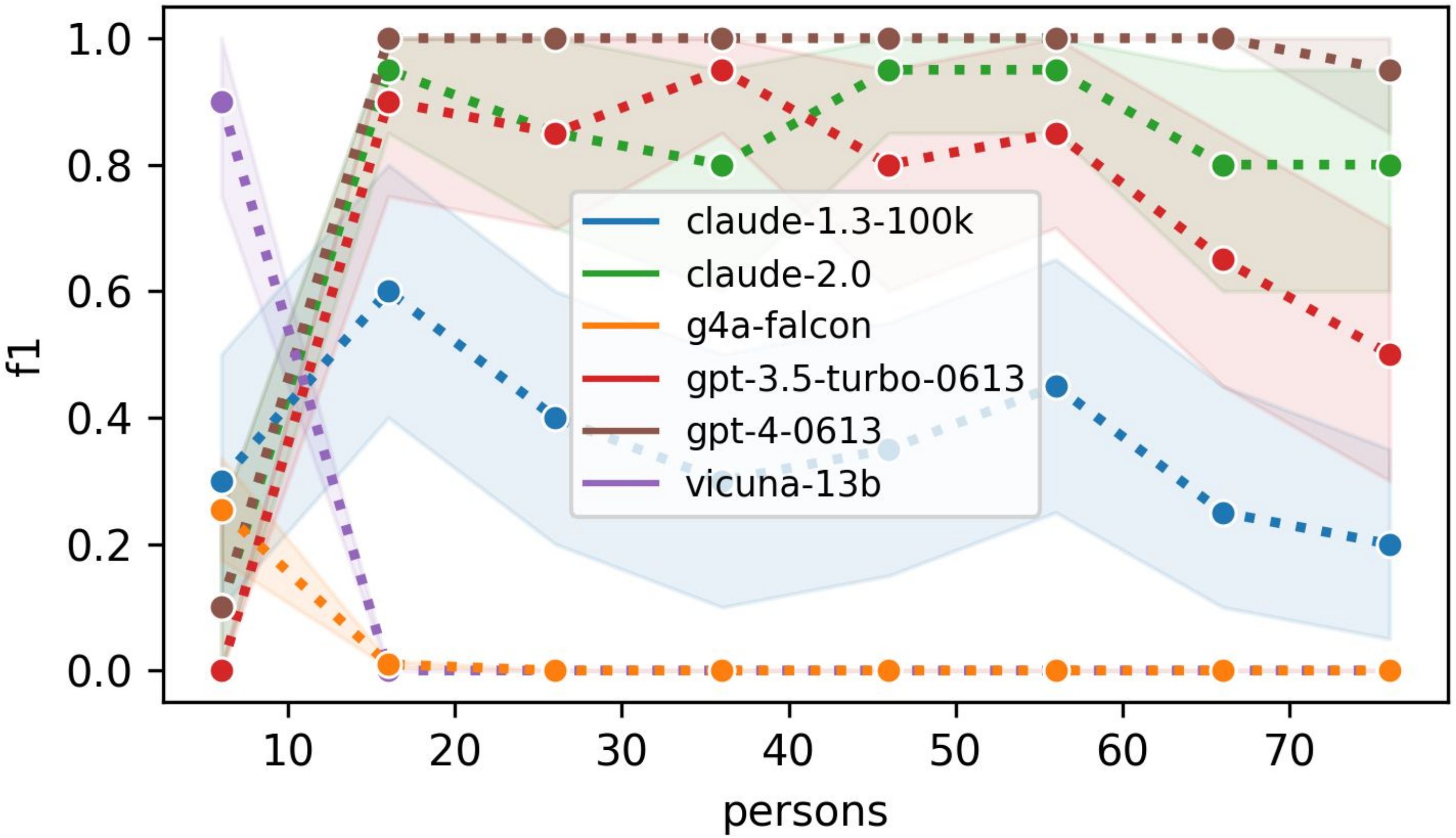
Task T3: Evaluation findings & mean of relative error

- Claude-1.3 misses prefix decl. or type statements (fixed in 2.0 □best)
- Ellipses lead to increased error rate for higher sizes (all sizes for Vicuna)
- Vicuna omits types for size 10 ??????
- Falcon lists prefixes only



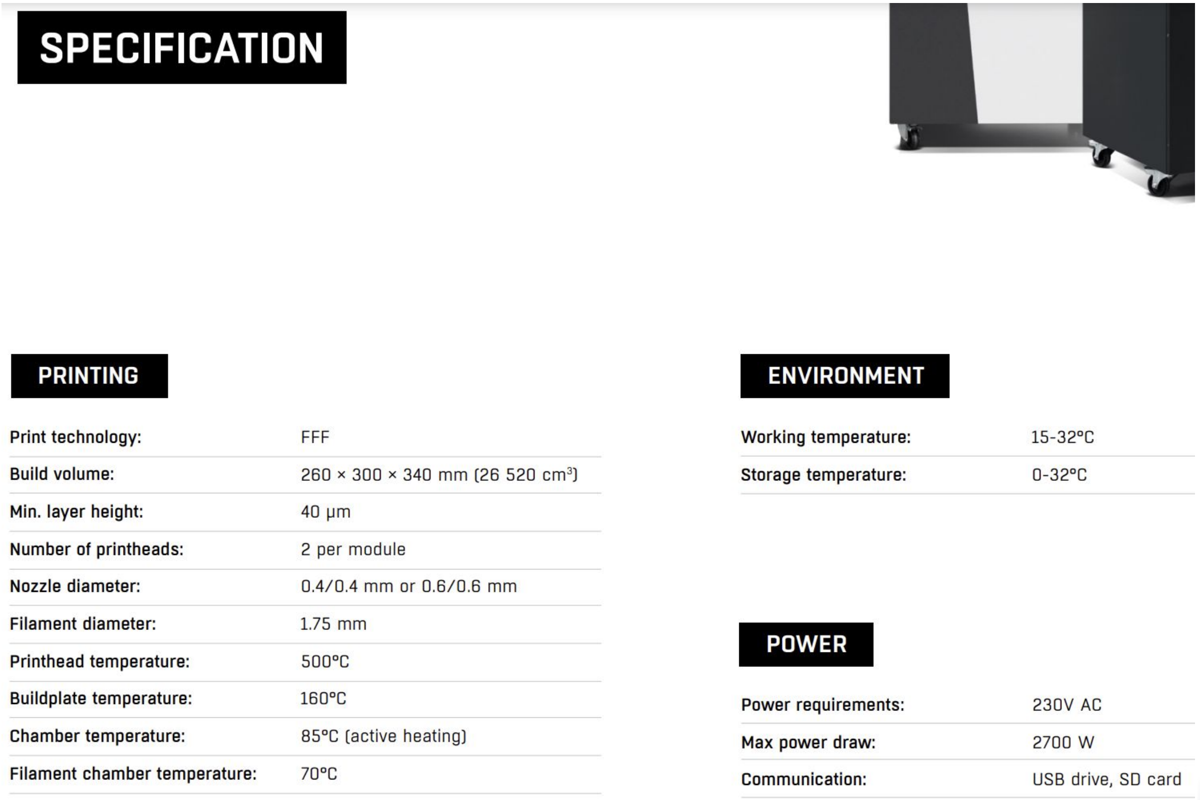
Task T4: Evaluation findings & F1 score

- GPT4 almost perfect
- Caude1.3/2.0/GPT3 sometimes confuse inlinks/outlinks
- Claude 1.3+GPT3.5 violate output constraint
- Vicuna/Falcon have incorrect reasoning, context window exceeded from 26/36 persons



“KISS” Task T5 – “Construct KG entity from Factsheet”

- **Construct an RDF entity based on textual key-value-style description**
- **Input is plaintext excerpt of one 3D printer PDF fact sheet**
- **Prompt/Actions very complex**
  - detailed w.r.t. how IRIs should look like (for clear comparison)
  - also challenges vocabulary knowledge
  - Heavy use of major prompt engineering techniques



SPECIFICATION

PRINTING

Print technology:	FFF
Build volume:	260 × 300 × 340 mm (26 520 cm^3)
Min. layer height:	40 µm
Number of printheads:	2 per module
Nozzle diameter:	0.4/0.4 mm or 0.6/0.6 mm
Filament diameter:	1.75 mm
Printhead temperature:	500°C
Buildplate temperature:	160°C
Chamber temperature:	85°C (active heating)
Filament chamber temperature:	70°C

ENVIRONMENT

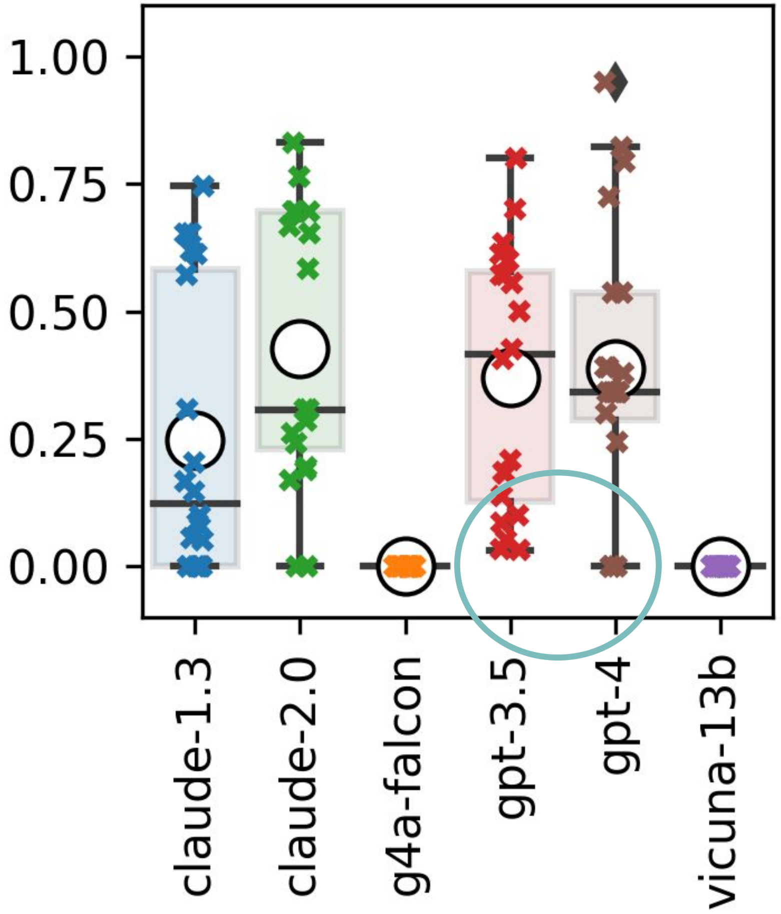
Working temperature:	15-32°C
Storage temperature:	0-32°C

POWER

Power requirements:	230V AC
Max power draw:	2700 W
Communication:	USB drive, SD card

Task T5: Evaluation findings & F1 score

- GPT4 has best almost perfect solution (outlier)
- GPT4 and Claude 1.3 frequently unparseable content, claude 2.0 sometimes too
- GPT3.5 no syntactic errors !!! !!!
- Falcon stuck in repetitive prefix gibberish patterns
- Vicuna – nothing looks like turtle


claude-1.3
claude-2.0
g4a-falcon
gpt-3.5
gpt-4
vicuna-13b