Can LLMs extract human-like fine-grained evidence for evidence-based fact-checking?

Antonín Jarolím, Martin Fajčík and Lucia Makaiová

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Motivation: Global Risks Report 2024-2025

"Please estimate the likely impact of the following risks over a 2-year period."

Motivation: Global Risks Report 2024-2025

"Please estimate the likely impact of the following risks over a 2-year period."

1th medium-term risk
(Elsner et al., 2025)

NATO releases revised AI strategy

10 July 2024

1 MIN READ

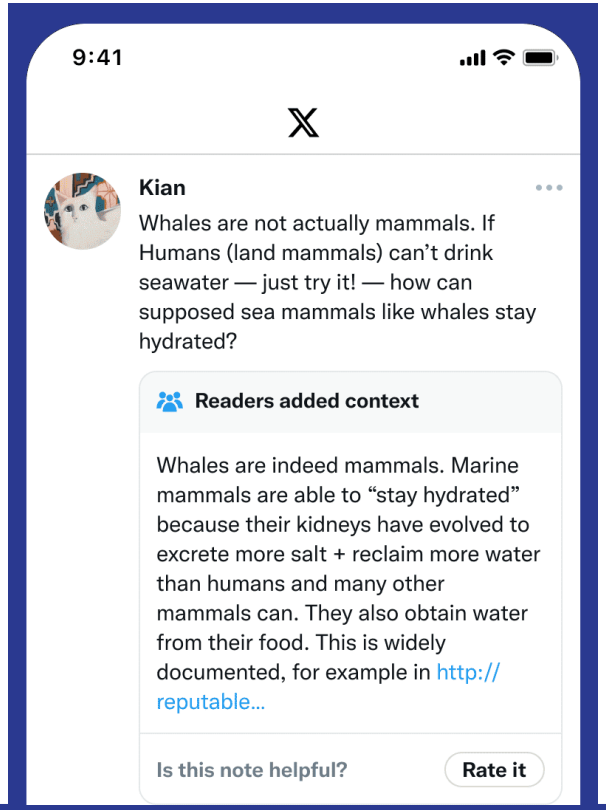
On Wednesday (10 July 2024), NATO released its revised artificial intelligence (AI) strategy, which aims to accelerate the use of AI technologies within NATO in a safe and responsible way.

It builds on one published in 2021 and takes account of recent advances in AI technologies, such as generative AI, and AI-enabled information tools.

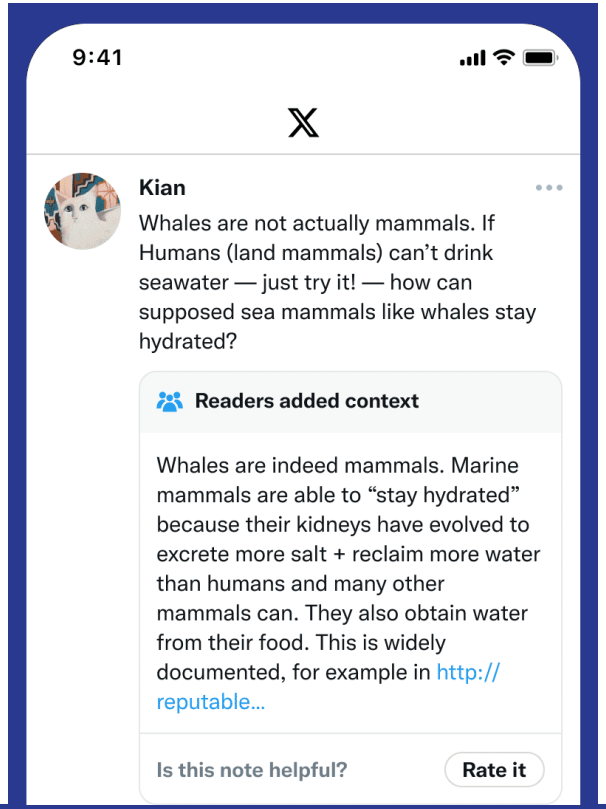
The strategy identifies several priorities, including: advancing the implementation of NATO's Principles of Responsible Use; increasing interoperability between AI systems throughout the Alliance; the combination of AI with other emerging disruptive technologies; and expanding NATO's AI ecosystem through closer cooperation with Allied industry and academia, NATO's Defence Innovation Accelerator DIANA, the NATO Innovation Fund and like-minded partners.

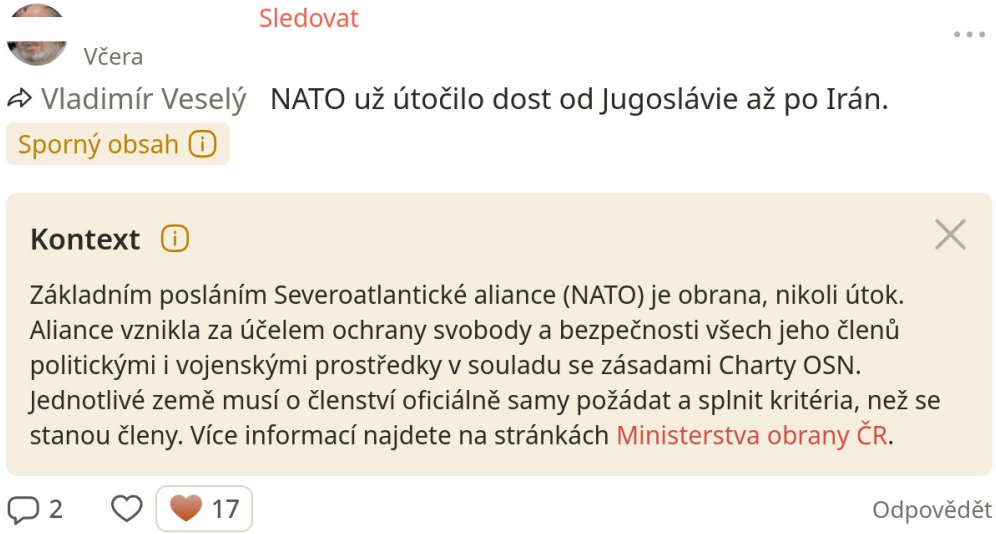
For the first time, the strategy identifies AI-enabled disinformation and information operations issues of concern for the Alliance, our societies and democracies.

Motivation: Platforms are already responding!



Motivation: Platforms are already responding!

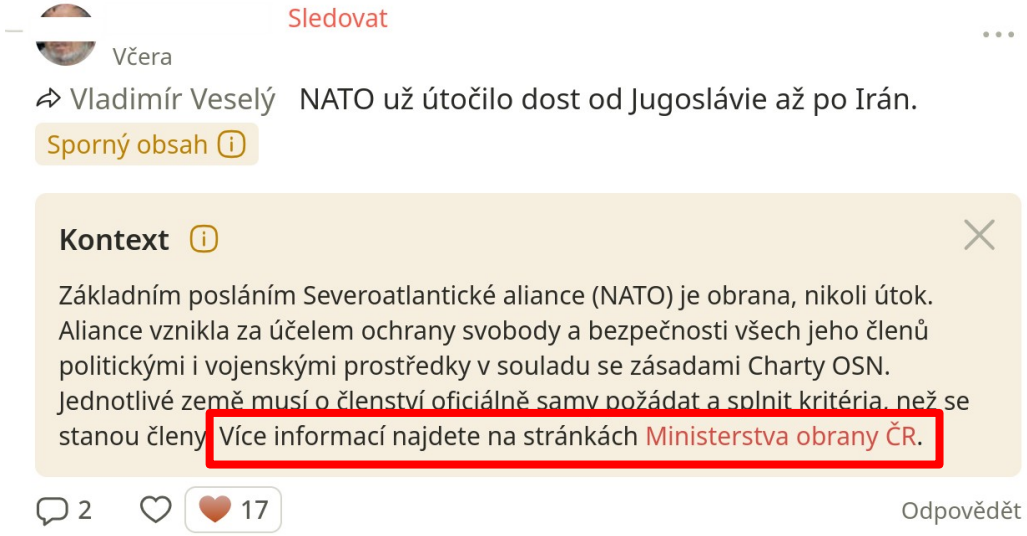




Motivation: Platforms are already responding!

Evidence:

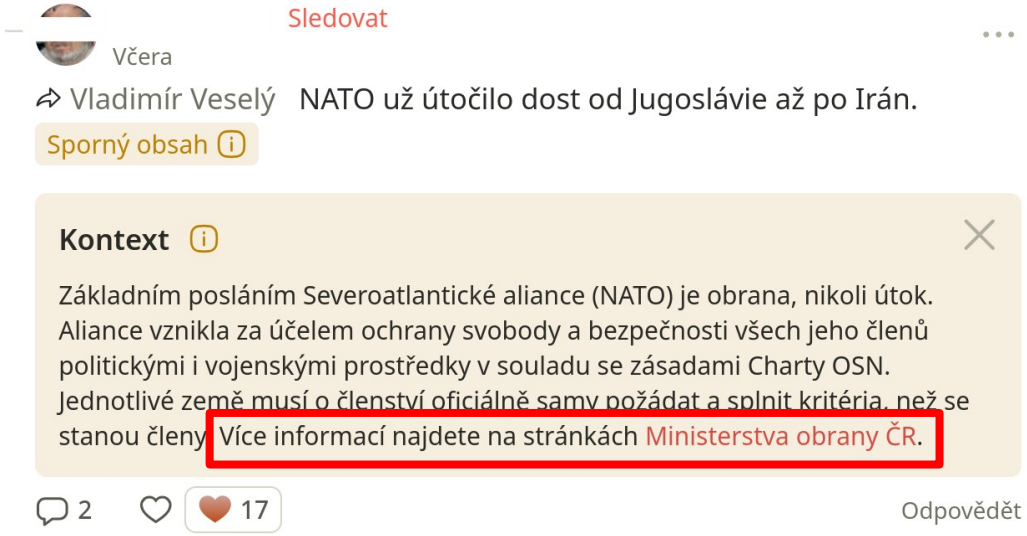
- increases user trust



Motivation: Platforms are already responding!

Evidence:

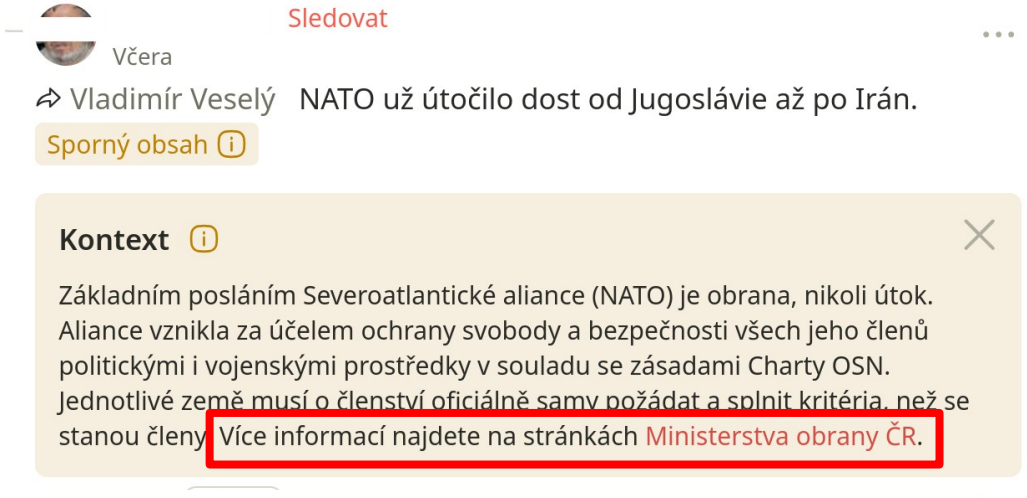
- increases user trust
- must be fine-grained



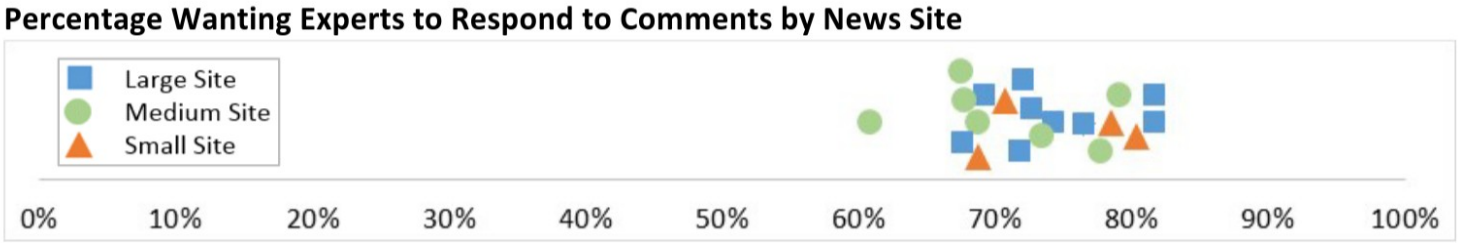
Motivation: Platforms are already responding!

Evidence:

- increases user trust
- must be fine-grained

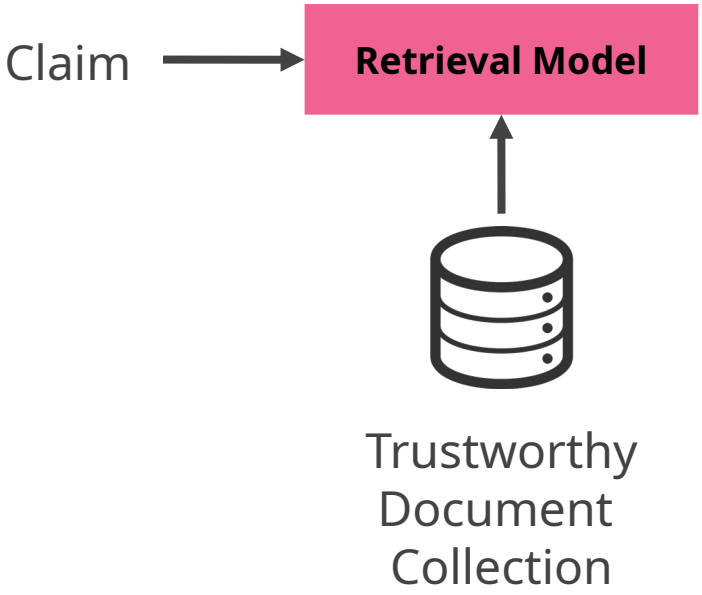


Moderation is requested!

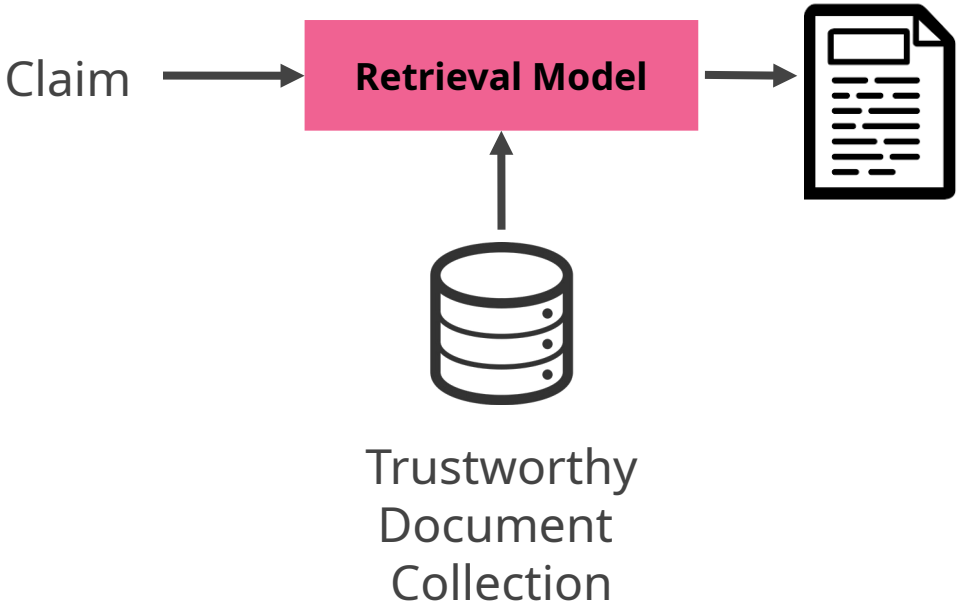


(Stroud et al., 2016)

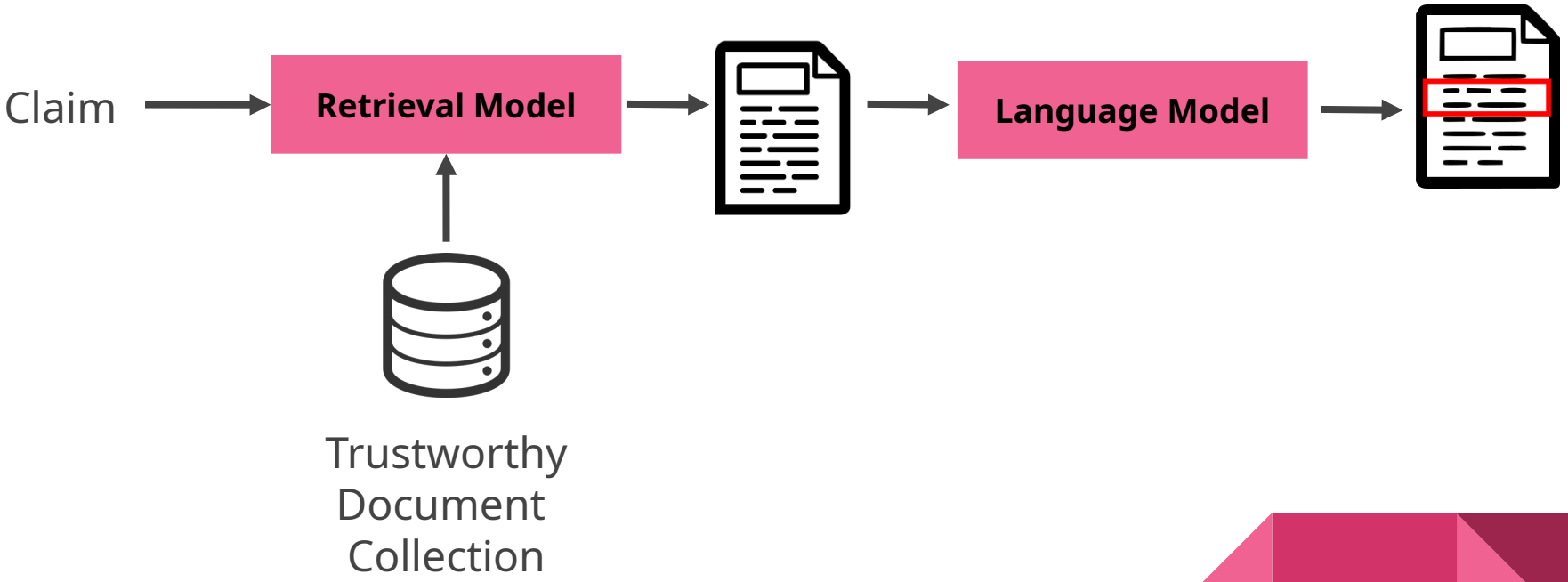
Automating Evidence-Based Fact Checking



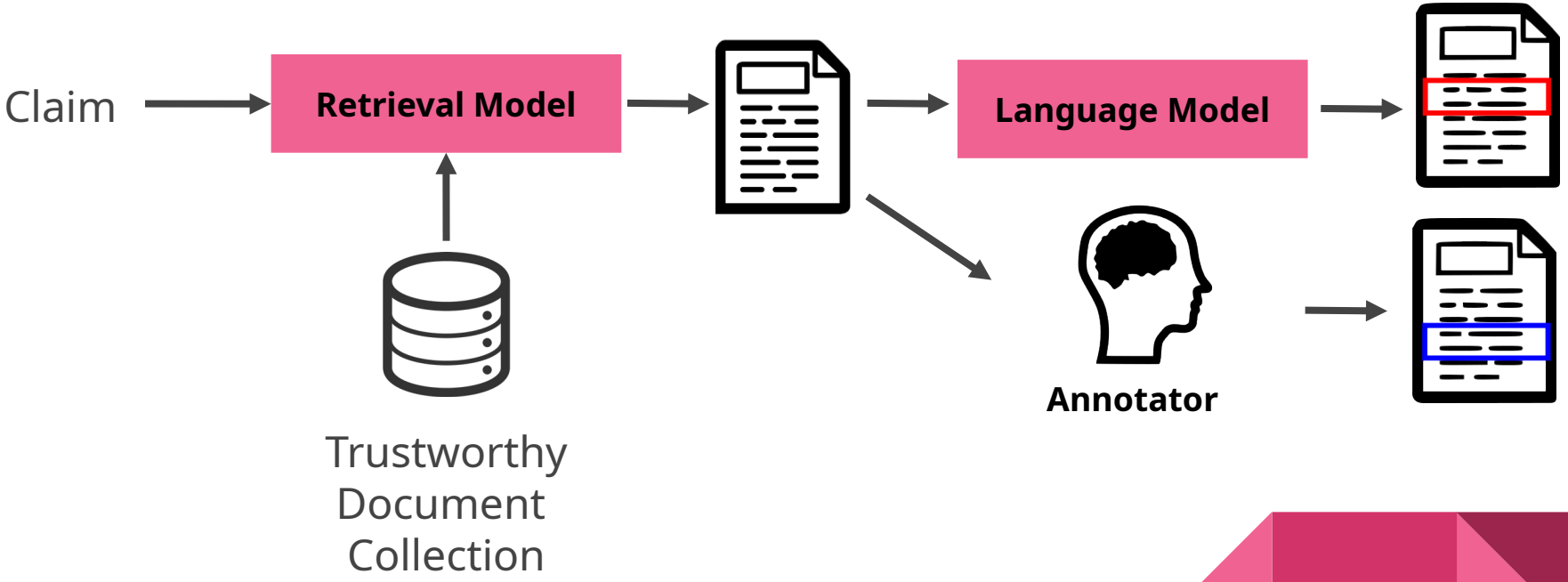
Automating Evidence-Based Fact Checking



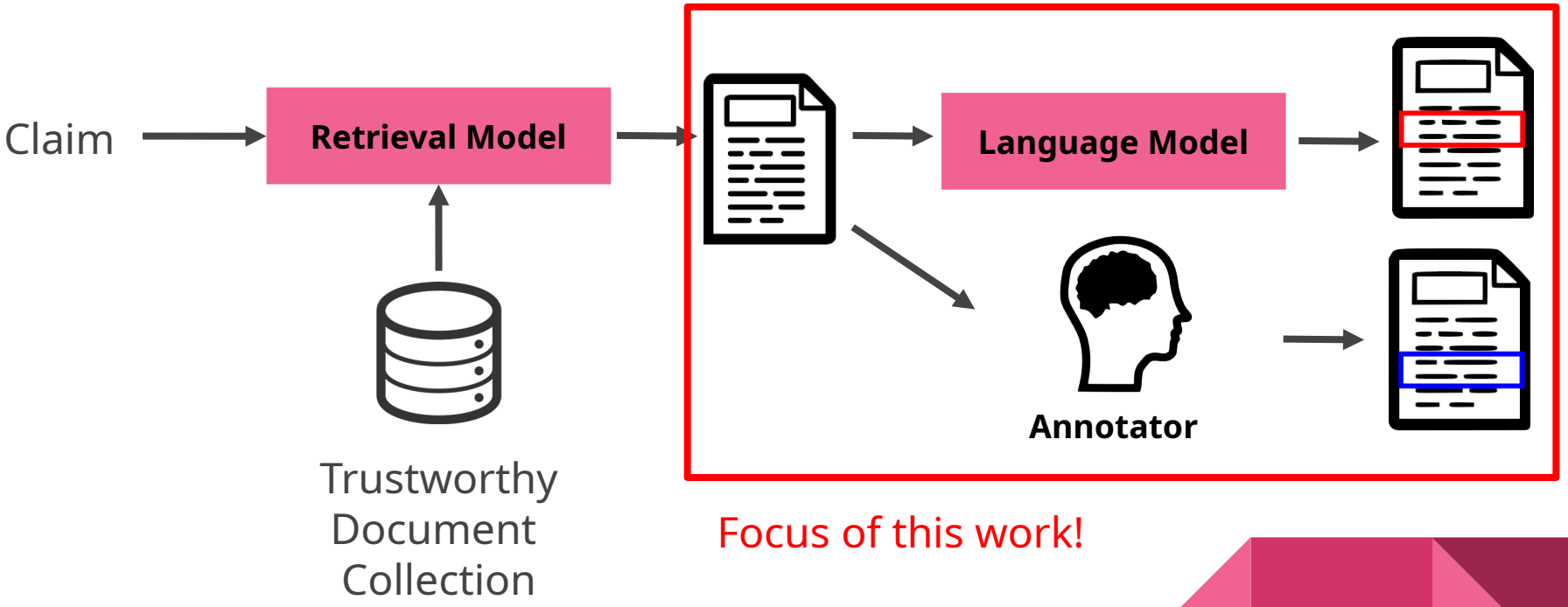
Automating Evidence-Based Fact Checking



Automating Evidence-Based Fact Checking



Automating Evidence-Based Fact Checking



Can LLMs extract human-like fine-grained evidence?

Can LLMs extract human-like fine-grained evidence?

1. Human Data Collection

Can LLMs extract human-like fine-grained evidence?

1. Human Data Collection
2. Inter-Annotator Agreement

Can LLMs extract human-like fine-grained evidence?

1. Human Data Collection
2. Inter-Annotator Agreement
3. Evidence Span Annotation using LLMs

Can LLMs extract human-like fine-grained evidence?

1. Human Data Collection
2. Inter-Annotator Agreement
3. Evidence Span Annotation using LLMs
4. LLM Annotation Performance

Human Fine-Grained Evidence Extraction

- No Czech or Slovak dataset provides fine-grained evidence for fact-checking.
- new dataset created:
  - 186 claim-document pairs
  - pool of 8 trained annotators
  - each pair annotated by 2 annotators

Human Fine-Grained Evidence Extraction

- No Czech or Slovak dataset provides fine-grained evidence for fact-checking.
- new dataset created:
  - 186 claim-document pairs
  - pool of 8 trained annotators
  - each pair annotated by 2 annotators

Guidelines:

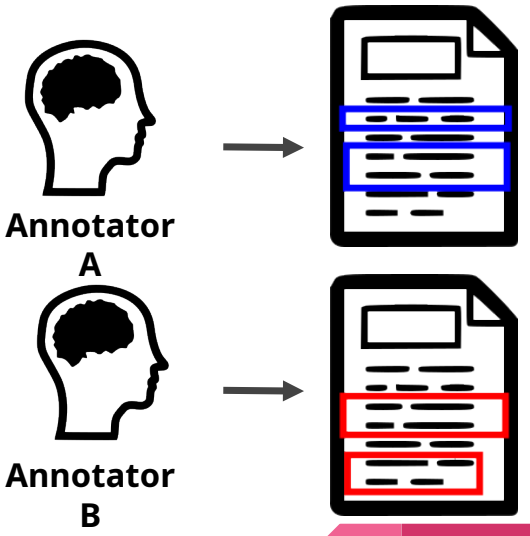
> Highlight the **minimal** portion of text that most convinces you that the given claim is **true**.

Human Fine-Grained Evidence Extraction

- No Czech or Slovak dataset provides fine-grained evidence for fact-checking.
- new dataset created:
  - 186 claim-document pairs
  - pool of 8 trained annotators
  - each pair annotated by 2 annotators

Guidelines:

> Highlight the **minimal** portion of text that most convinces you that the given claim is **true**.



Assessing Inter-Annotator Agreement

Assessing Inter-Annotator Agreement

- multiple evidence span annotations

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans
- -> Hungarian matching

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans
- -> Hungarian matching

Evidence Span 1
Evidence Span 2
Evidence Span 3

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans
- -> Hungarian matching

Evidence Span 1
Evidence Span 2
Evidence Span 3

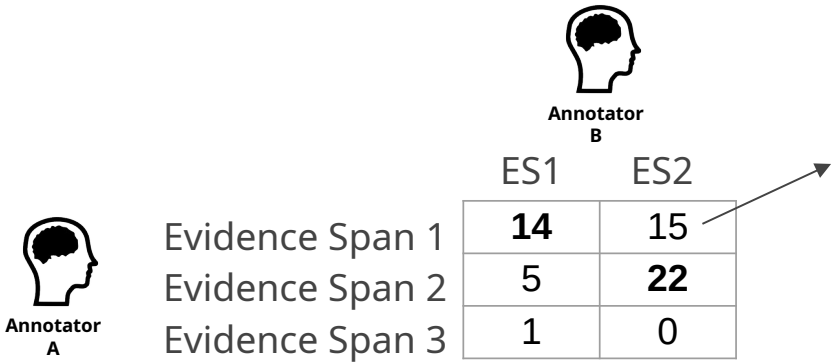
Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans
- -> Hungarian matching

	ES1	ES2
Evidence Span 1	**14**	15
Evidence Span 2	5	**22**
Evidence Span 3	1	0

Assessing Inter-Annotator Agreement

- multiple evidence span annotations
- different annotation exhaustivity
- do not penalize different number of spans
- -> Hungarian matching



F1-Score
- word level
- removed stop-words

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Claim: *The Call of Duty game was successful.*

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.

Evidence annotation:

```
{
  "start": 42,
  "end": 118,
  "label": "supporting"
}
```

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Evidence annotation:

```
{
  "start": 42,
  "end": 118,
  "label": "supporting"
}
```

Extracting Annotations from LLM Outputs

1. Listing start and end indices of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Evidence annotation:

```
{
  "start": 42,
  "end": 118,
  "label": "supporting"
}
```

❌ “GPT-4 frequently generated indices not matching the corresponding spans in a paragraph” (Hasanain et al.)

Extracting Annotations from LLM Outputs

Extracting Annotations from LLM Outputs

2. Labeling the text with special tokens

Extracting Annotations from LLM Outputs

2. Labeling the text with special tokens

Claim: *The Call of Duty game was successful.*

Extracting Annotations from LLM Outputs

2. Labeling the text with special tokens

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.

Extracting Annotations from LLM Outputs

2. Labeling the text with special tokens

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.

Evidence annotation:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***<start_span>became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue<end_span>****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Extracting Annotations from LLM Outputs

2. Labeling the text with special tokens

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.

Evidence annotation:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***<start_span>became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue<end_span>****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

❌ regeneration of entire text is required

Extracting Annotations from LLM Outputs

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Claim: *The Call of Duty game was successful.*

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Evidence annotation:
{
"text": [“**became the fastest-selling Call of Duty title to reach 1 billion dollar in revenue”**],
"label": "supporting"
}

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Evidence annotation:

```
{
  "text": [“became the fastest-selling Call of Duty title to reach 1 billion dollar in revenue”],
  "label": "supporting"
}
```

✅ does not need to regenerate entire text
❌ must copy text spans

Extracting Annotations from LLM Outputs

3. Listing textual content of spans

Claim: *The Call of Duty game was successful.*

Input text:
Game received generally positive reviews from critics. It was a commercial success and broke several franchise records, including that it ***became the fastest-selling Call of Duty title to reach 1 billion dollars in revenue****. A sequel titled Call of Duty: Modern Warfare III was released on 10 November 2023.*

Evidence annotation:

```
{
  "text": [“became the fastest-selling Call of Duty title to reach 1 billion dollar in revenue”],
  "label": "supporting"
}
```

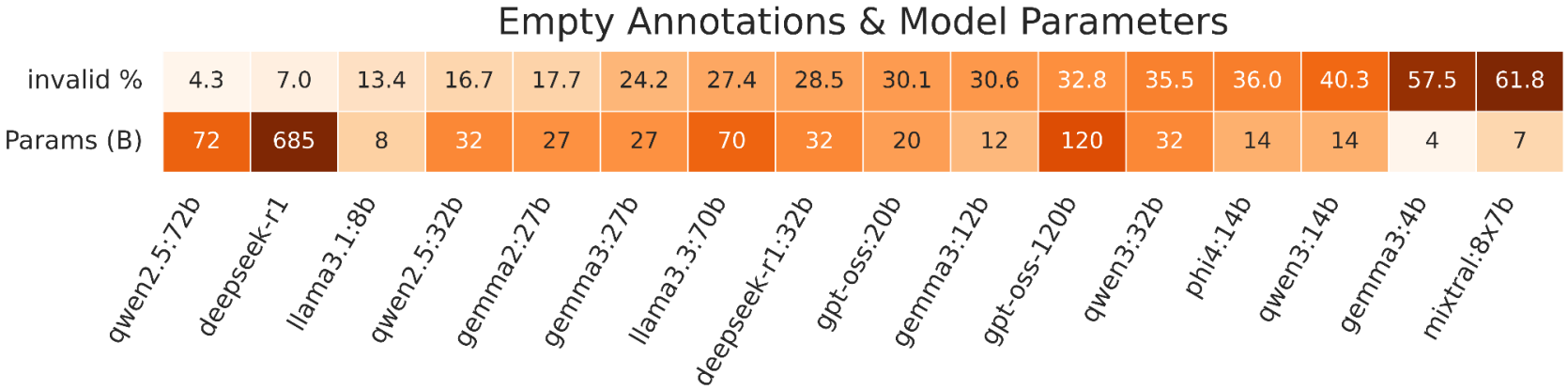
✅ does not need to regenerate entire text
❌ must copy text spans

Experiment 1: Number of Evidence Copy Errors

- at least one span not present verbatim -> invalid

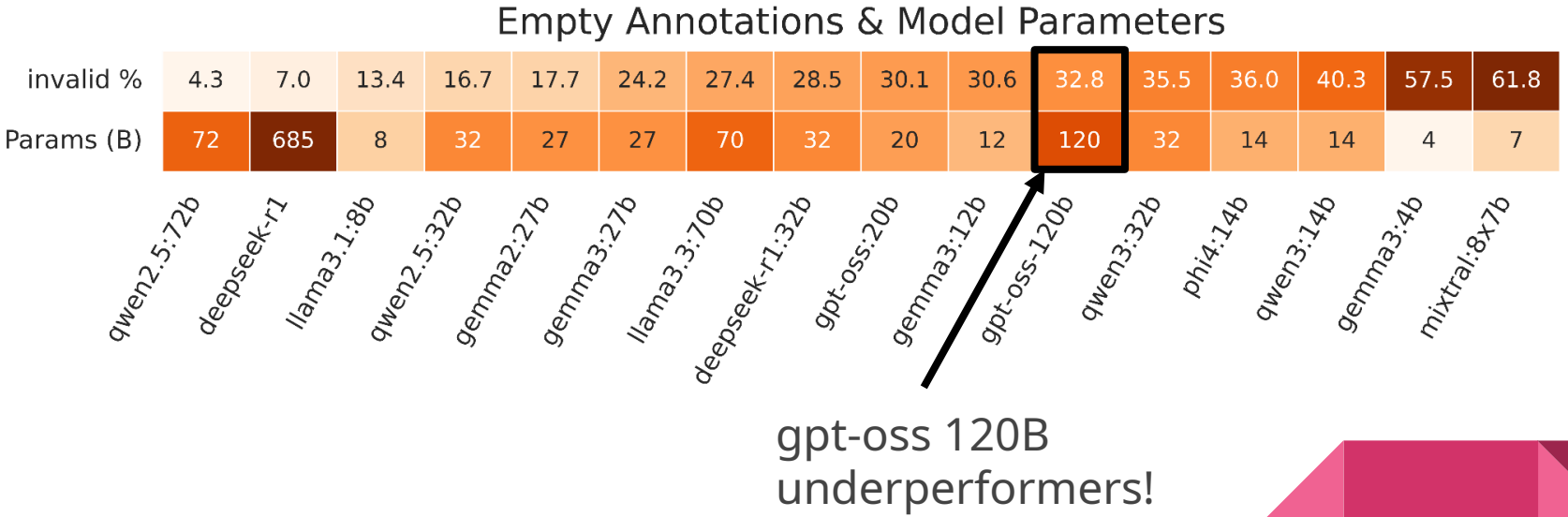
Experiment 1: Number of Evidence Copy Errors

- at least one span not present verbatim -> invalid



Experiment 1: Number of Evidence Copy Errors

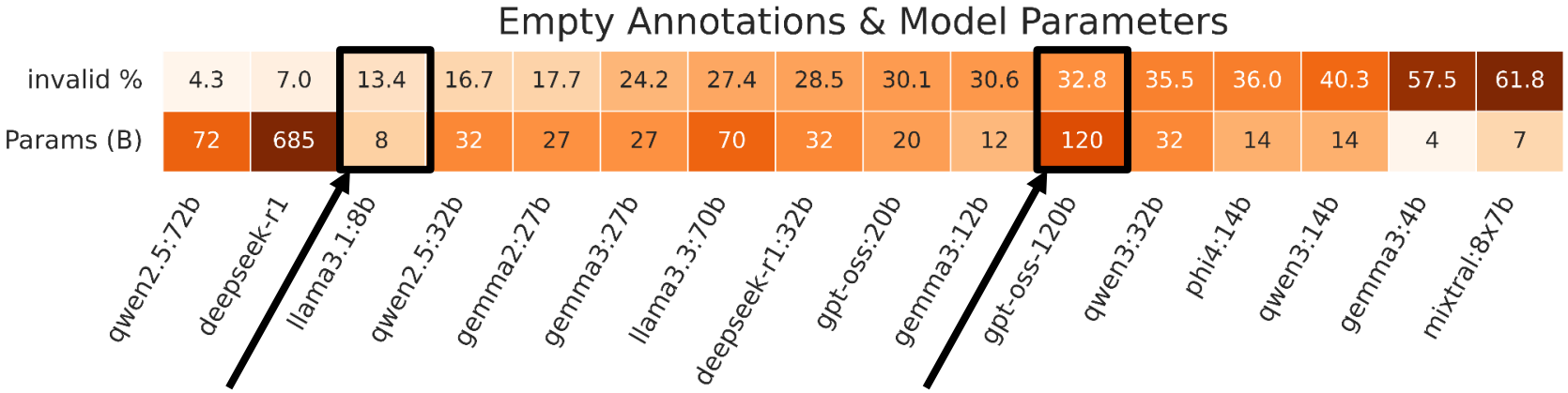
- at least one span not present verbatim -> invalid



gpt-oss 120B underperformers!

Experiment 1: Number of Evidence Copy Errors

- at least one span not present verbatim -> invalid

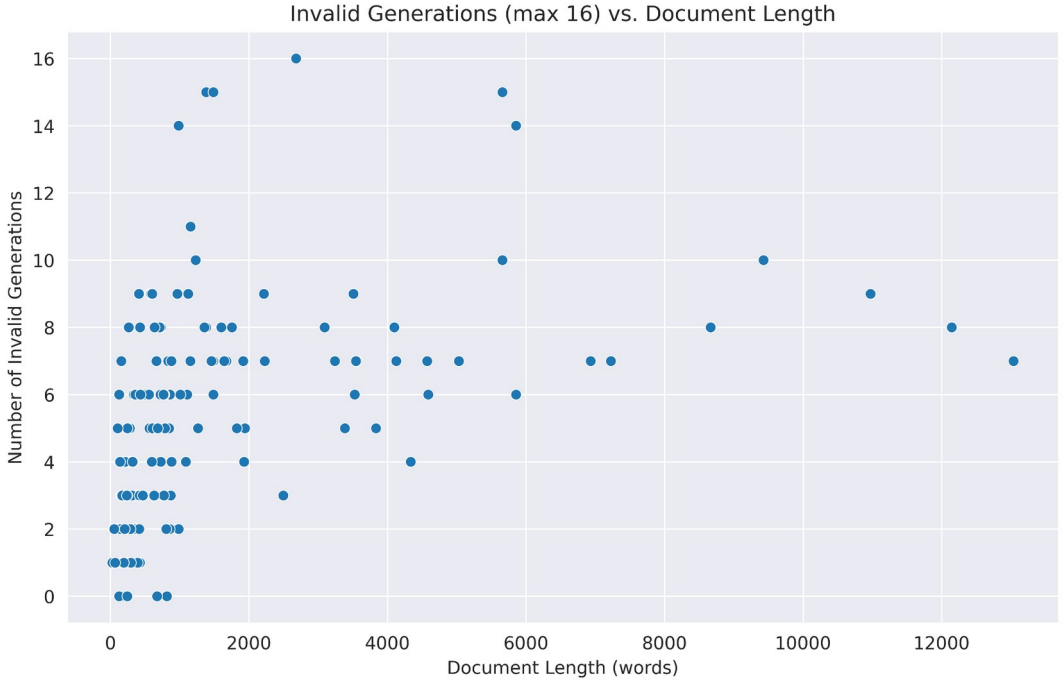


llama3.1 8B
exceeds expectations!

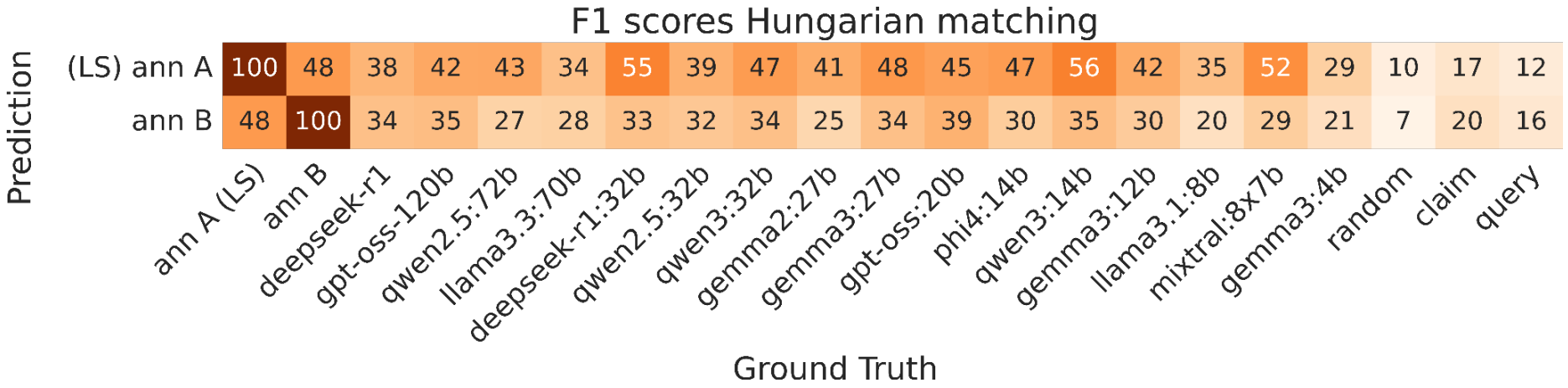
gpt-oss 120B
underperformers!

Experiment 1: Number of Evidence Copy Errors

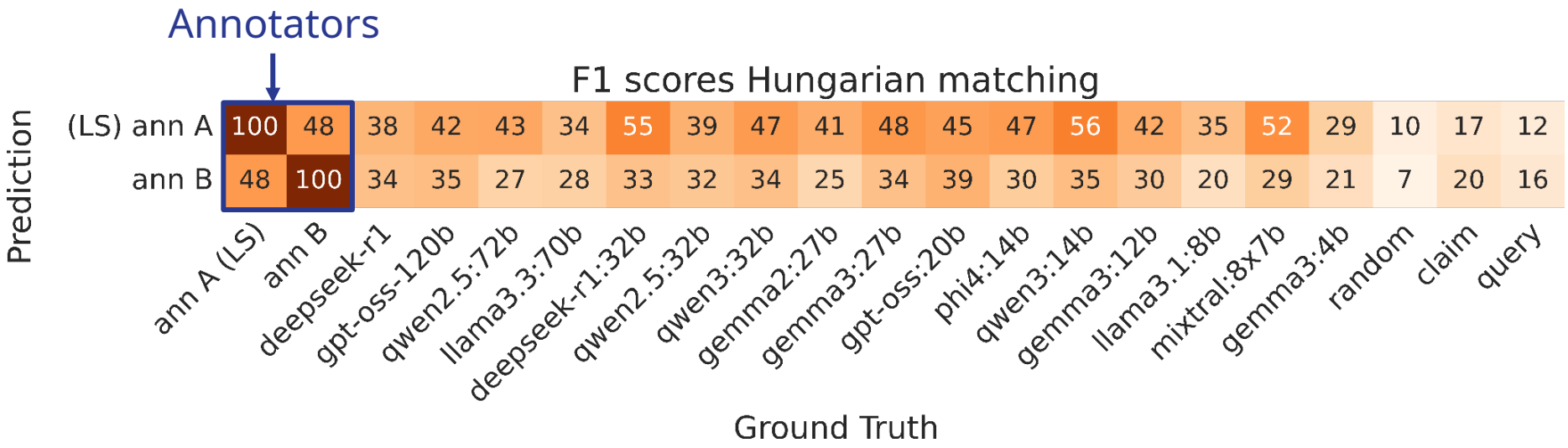
- does the size matter?



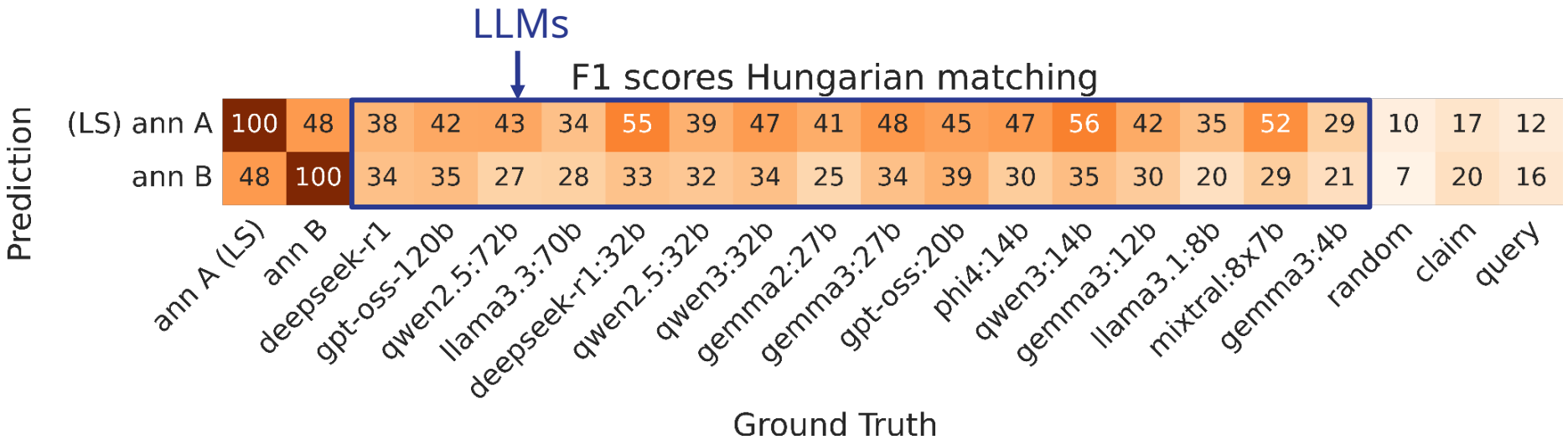
Experiment 2: Human – LLM scores



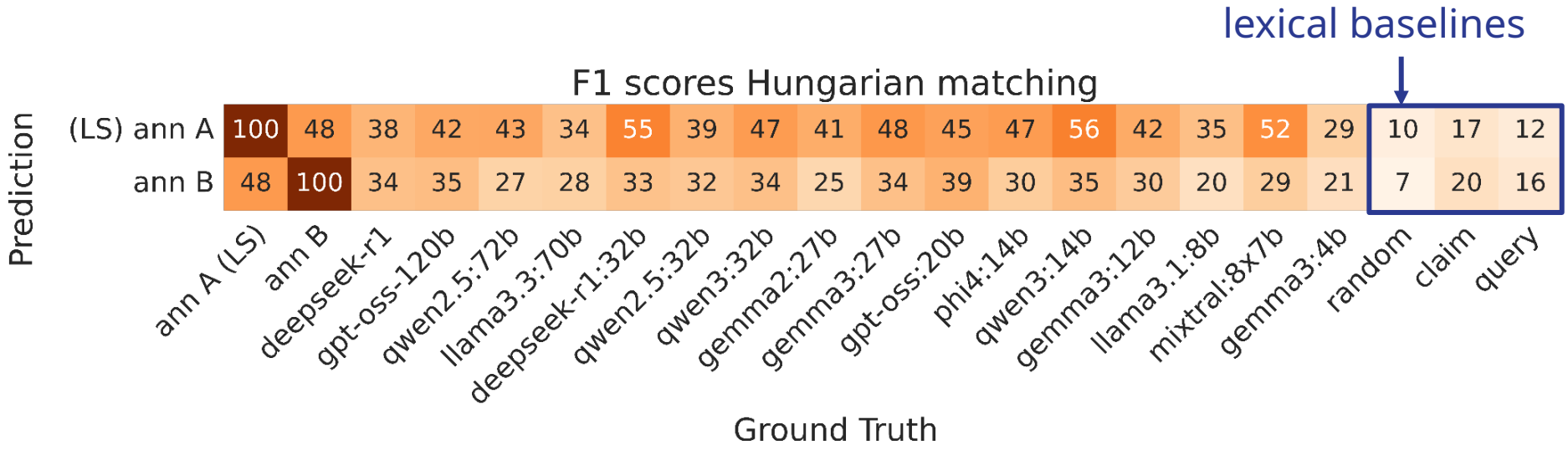
Experiment 2: Human – LLM scores



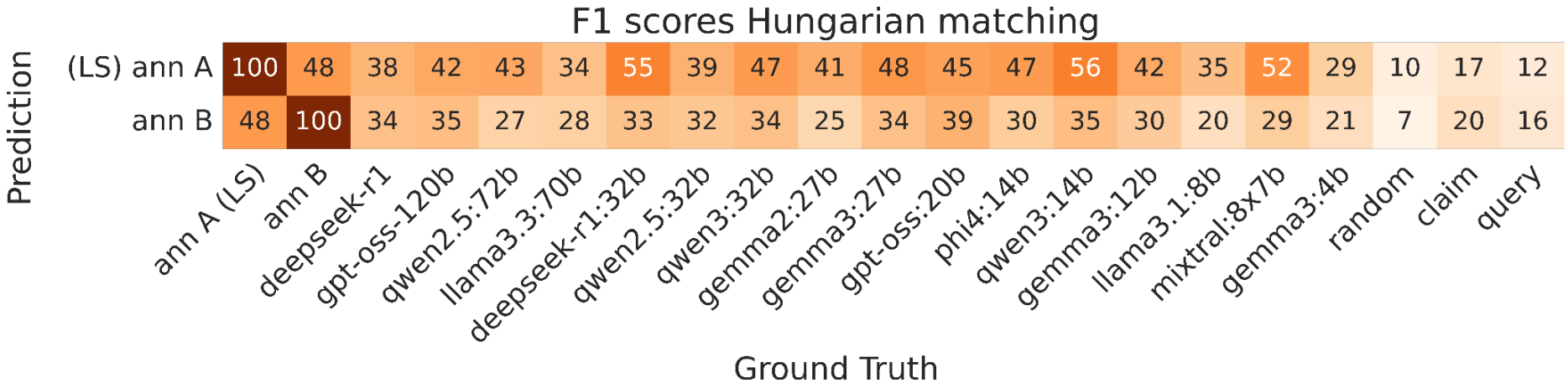
Experiment 2: Human – LLM scores



Experiment 2: Human – LLM scores

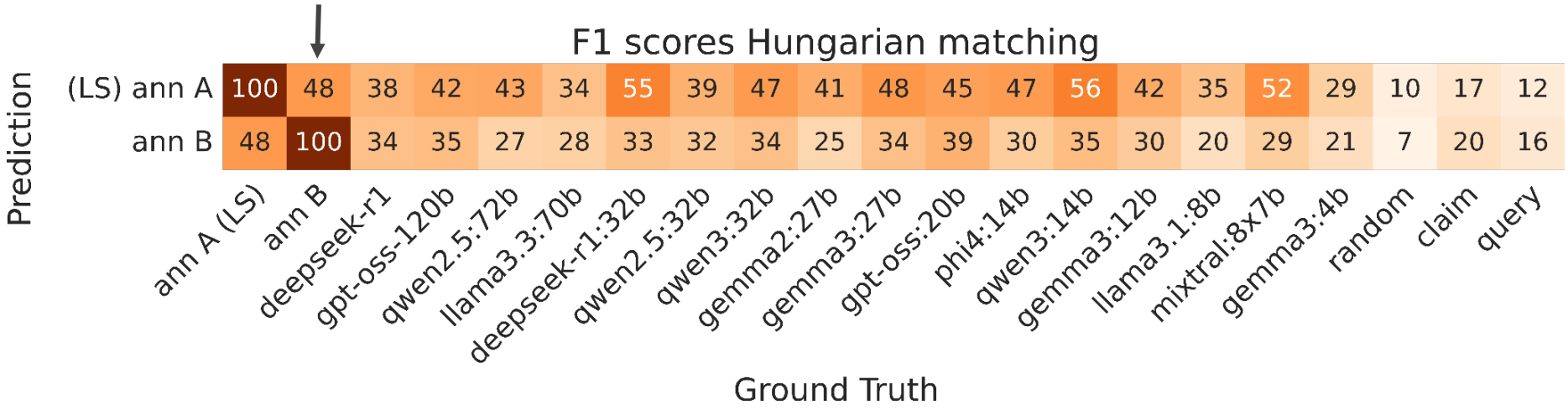


Experiment 2: Human – LLM scores



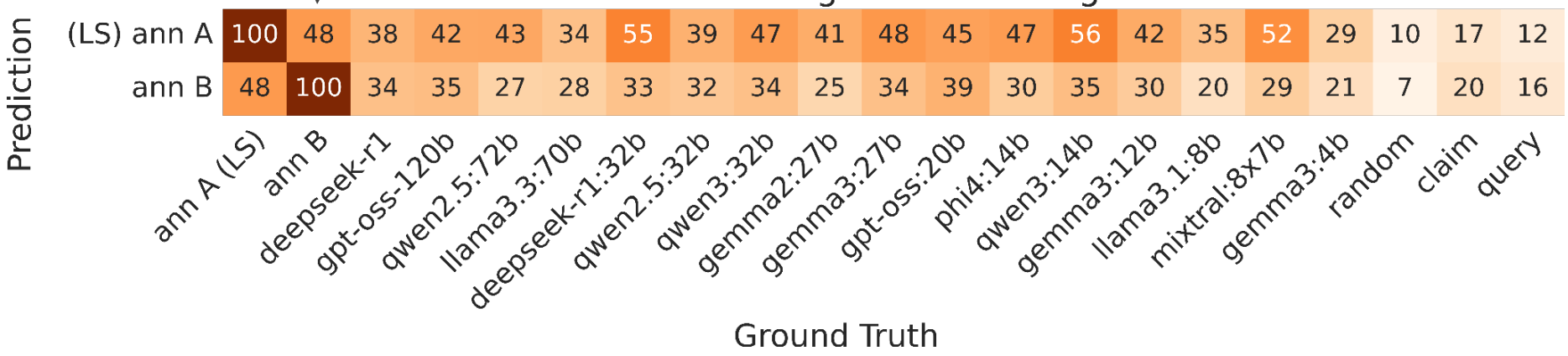
Experiment 2: Human – LLM scores

1. IAA – Higher than MOST LLMs



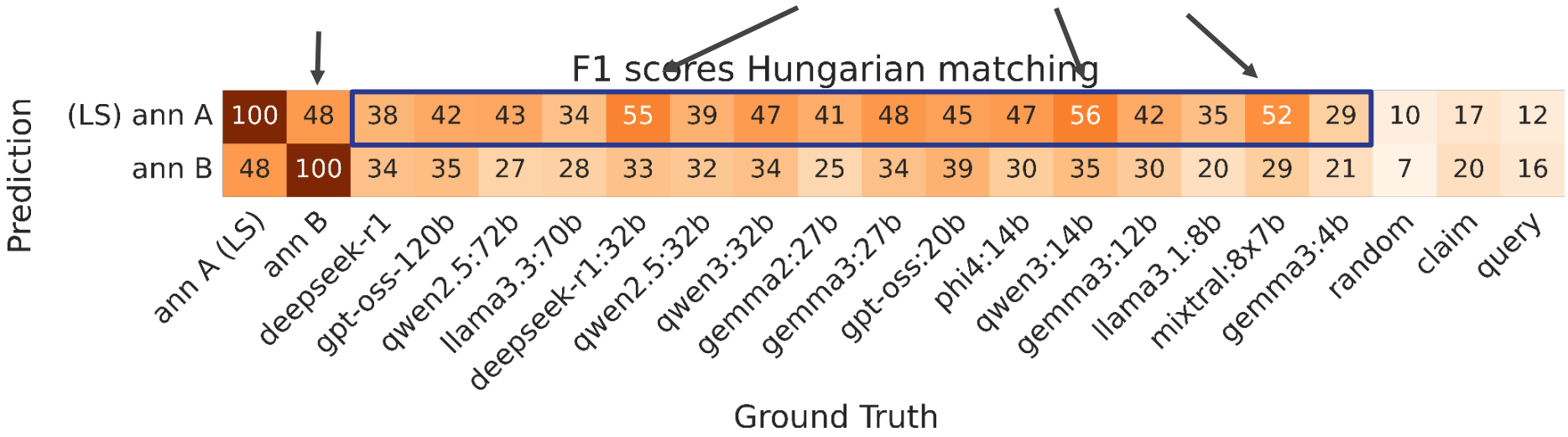
Experiment 2: Human – LLM scores

1. IAA – Higher than MOST LLMs
2. Sometimes even higher !?



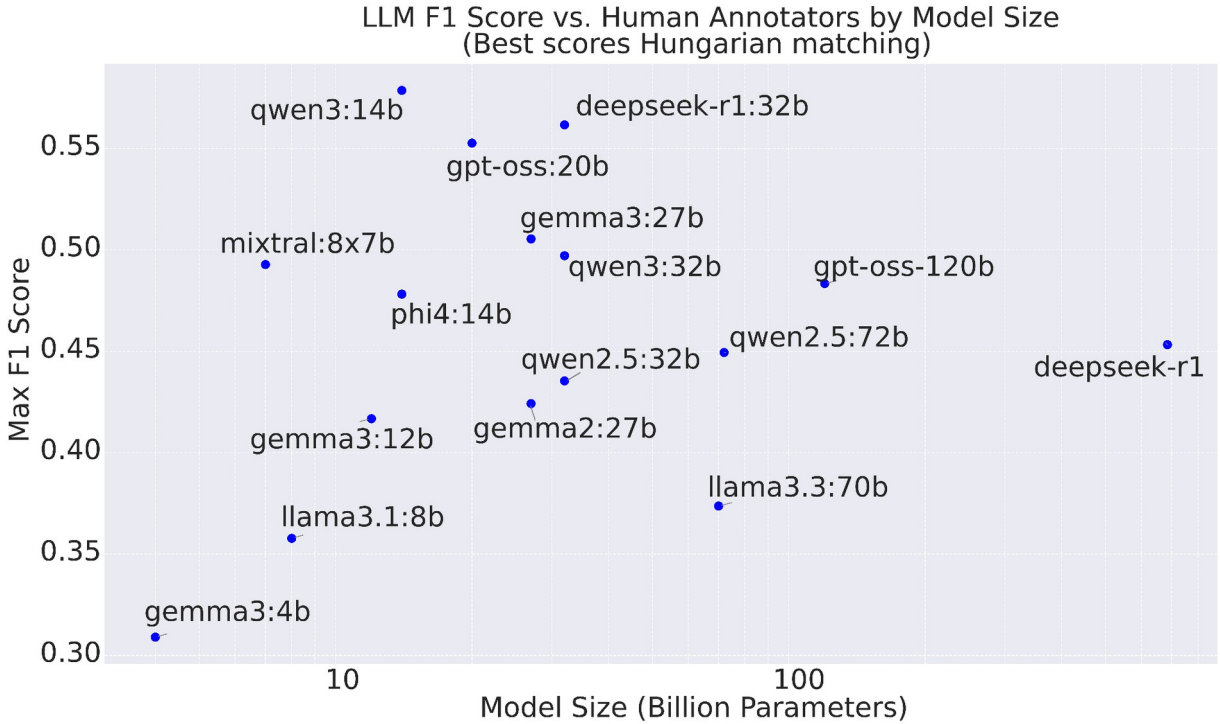
Experiment 2: Human – LLM scores

1. IAA – Higher than MOST LLMs
2. Sometimes even higher !?

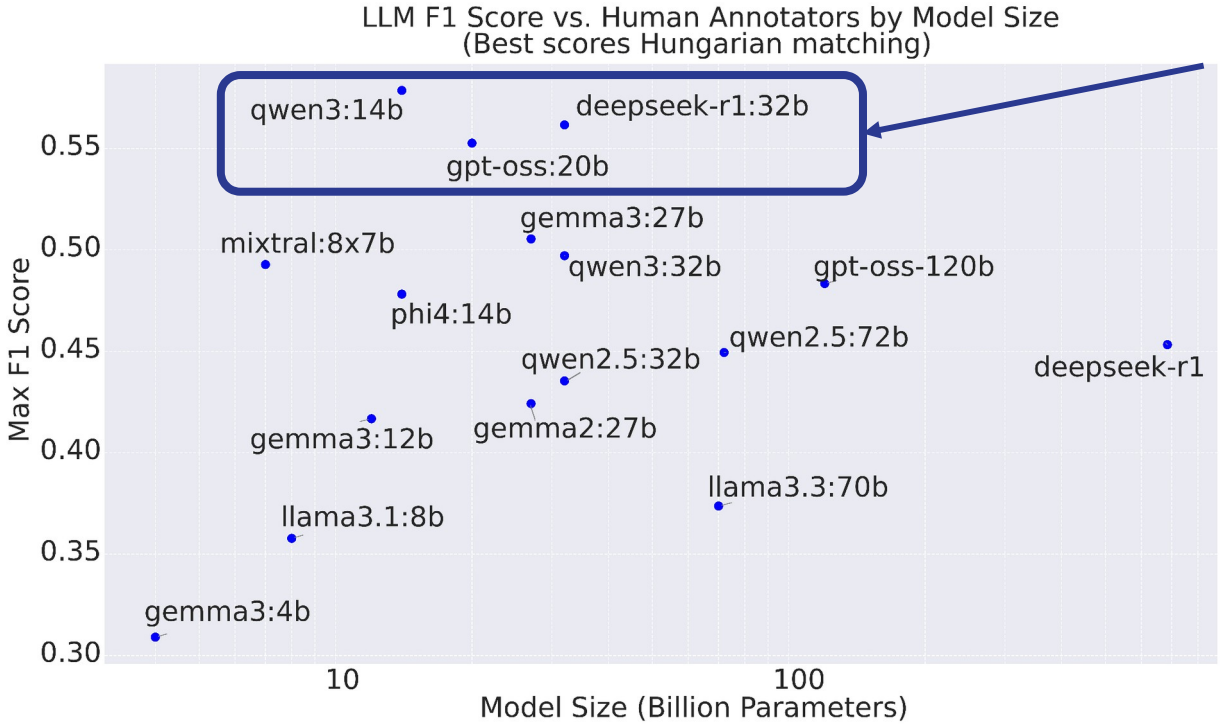


3. Consistently better match with Label Studio Annotations

Experiment 3: LLM scores vs Model size

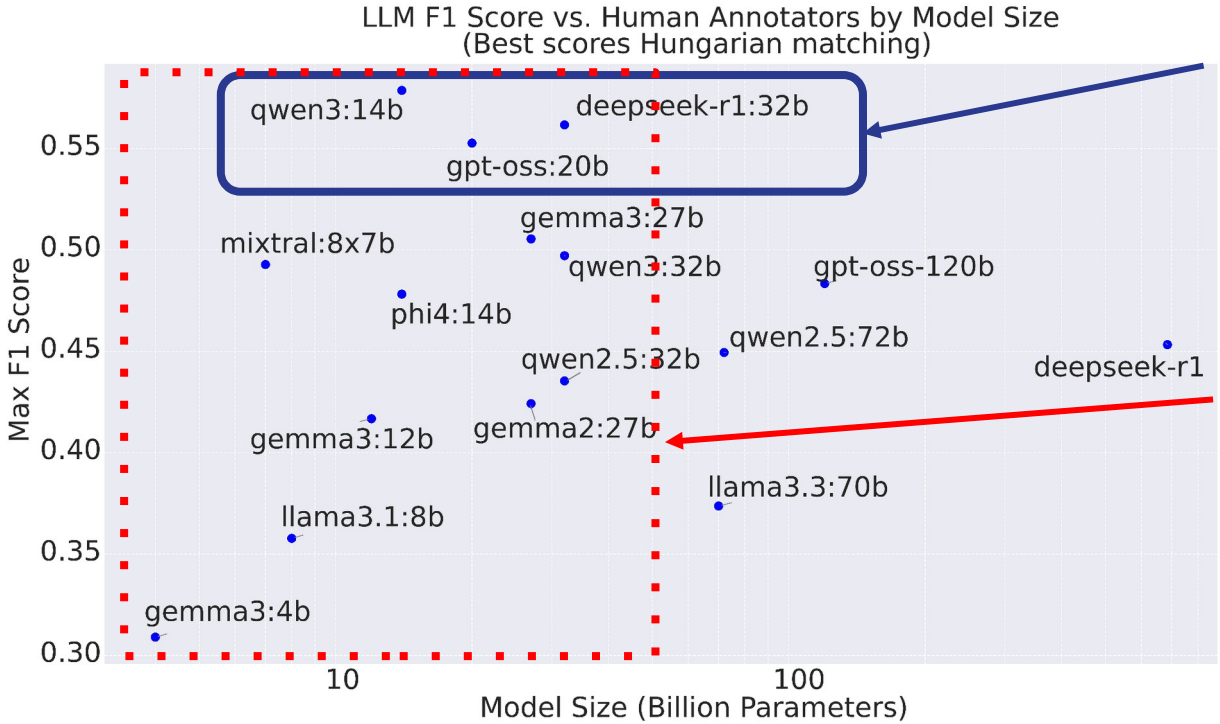


Experiment 3: LLM scores vs Model size



1. optimal trade-off

Experiment 3: LLM scores vs Model size



1. optimal trade-off

2. performance increases only up to a threshold

Improving LLM Regeneration Mechanism (Copy)

1. Custom Decoding Algorithm
- set logits of invalid tokens to -inf

2. Grammar-Constrained Decoding
- input constrained (Geng et al., 2023)

3. LLM with Explicit Copy Mechanism
- CopyLlama3.2-3B
- ongoing work (Fajčík et al., 2025)

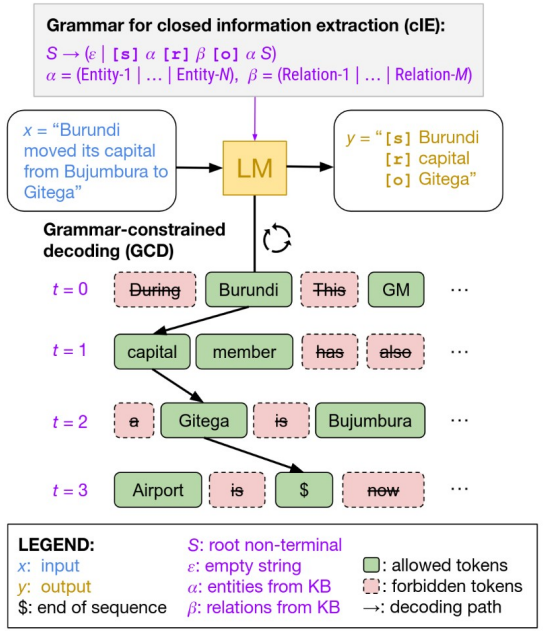


Figure 1: **Grammar-constrained decoding (GCD),** applied to the task of closed information extraction, where

Conclusions

- LLMs fail to copy spans from context verbatim
- increasing challenge on
  - Czech / Slovak data
  - long documents
- evidence extraction increases only up to a model size threshold
- best trade-off between model size and alignment:
  - gwen 3 (14B)
  - deepseek-r1 (32B)
  - gpt-oss (20B)

References

1. Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. xcomet : Transparent machine translation evaluation through fine-grained error detection. Trans. Assoc. Comput. Linguistics, 12:979–995, 2024. doi: 10.1162/TACL_A_00683. URL https://doi.org/10.1162/tacl_a_00683.
2. Mark Elsner, Grace Atkinson, and Saadia Zahidi. Global risks report 2025. Technical report, World Economic Forum, Geneva, January 2025. 20th edition; based on the Global Risks Perception Survey 2024–2025.
3. Geng, Saibo, et al. "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.
4. Stroud, N.J., Van Duyn, E., Alizor, A., Alibhai, A., Lang, C.: Comment section survey across 20 news sites. Report published by the Center for Media Engagement, University of Texas at Austin (2017)