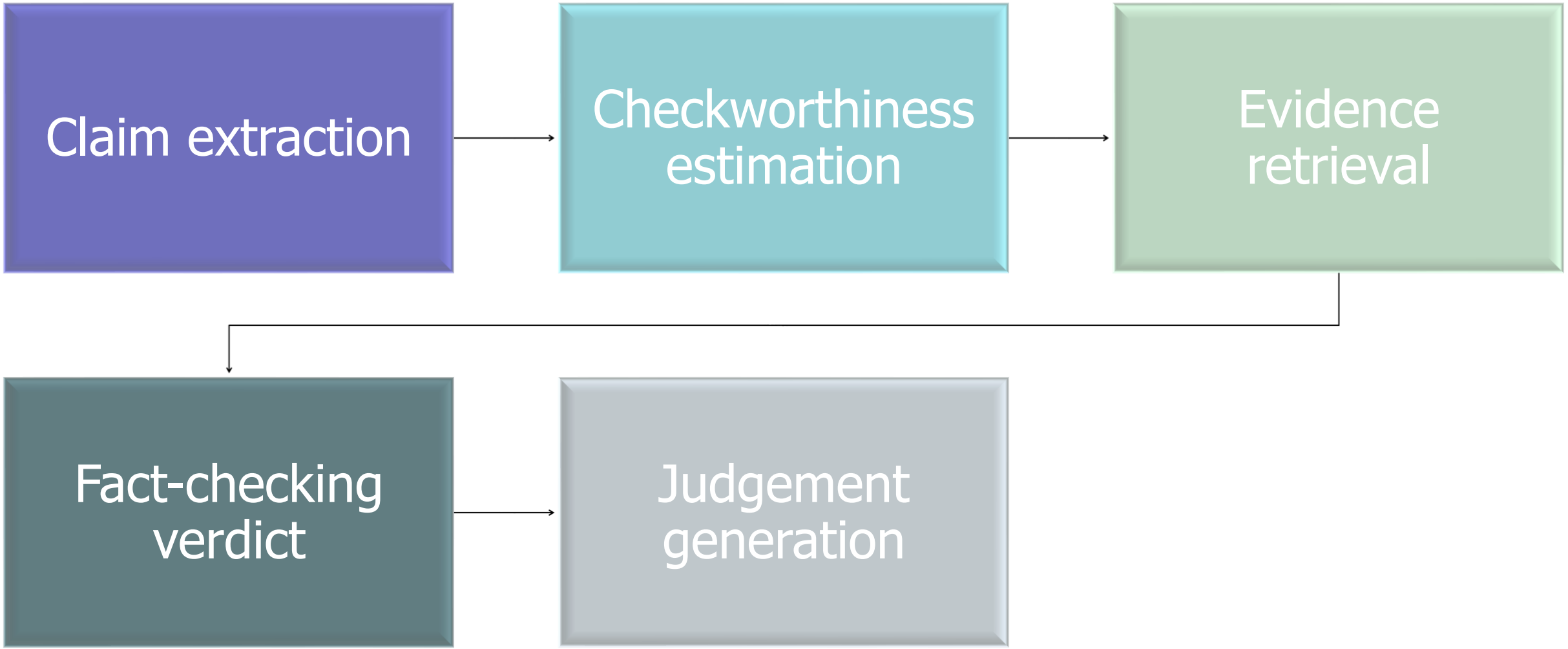
# Examining the Metrics for Document-Level Claim Extraction in Czech and Slovak

Lucia Makaiová, Martin Fajčík and Antonín Jarolím

# Fact checking pipeline



# CLAIM EXTRACTION DATASET

**GCD (Czech Gold Claims Dataset)**

- 60 comments
- 2.17 claims per comment on average
- Czech only data from single annotator

**GCD (Czech Gold Claims Dataset)**

- 60 comments
- 2.17 claims per comment on average
- Czech only data from single annotator

**CSAD (Czech-Slovak Agreement Dataset)**

- 120 comments
- 3.12 claims per comment on average
- Czech and Slovak data from 9 annotators

**GCD (Czech Gold Claims Dataset)**

- 60 comments
- 2.17 claims per comment on average
- Czech only data from single annotator

**CSAD (Czech-Slovak Agreement Dataset)**

- 120 comments
- 3.12 claims per comment on average
- Czech and Slovak data from 9 annotators

**Pekny a trefny clanek. Osobne vyznavam tvrdy pristup jaky maji v Oregonu. Tedy extremni pokuty a postihy za znecistovani prirody a nerespektovani narizeni. Ve zminenem Oregonu treba pokuta za znecistovani az do $6250.**

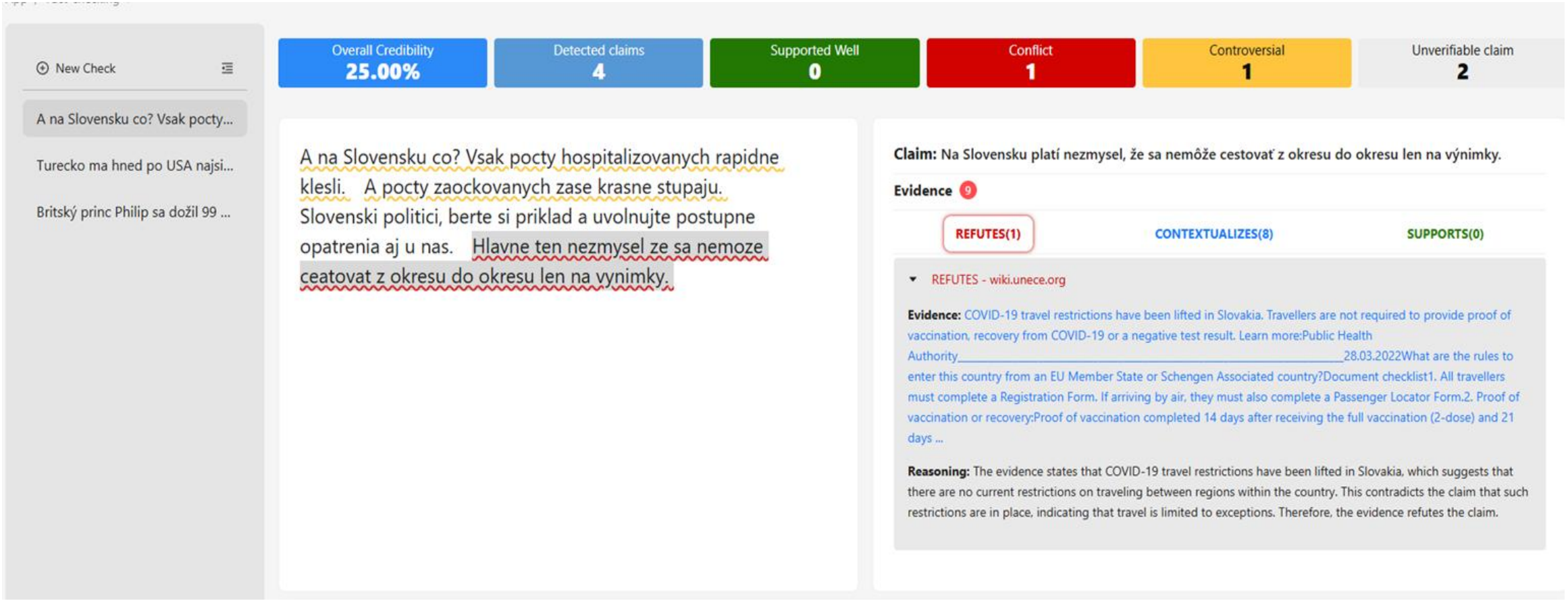
**V Oregone majú extrémne pokuty a postihy za znečisťovanie prírody a nerešpektovanie nariadení.**

**V Oregone sa pokuta za znečisťovanie môže vyšplhať až do 6250 dolárov.**

# CLAIM EXTRACTION SYSTEM

# Claim extraction -system

- automatic tool for claim extraction to assist our annotators
- LOKI fact checking pipeline works (well) even on czech data




Loki: An Open-Source Tool for Fact Verification (Li et al., COLING 2025)

# Claim extraction -system

Models for extraction: GPT 4omini, GPT 4.1, Claude Sonnet 3.7

Attributes of claims:

- **atomic** = expresses a single factual proposition
- **decontextualized** = self-contained and interpretable without relying on the surrounding text
- **check-worthy** = factual claim, general public will be interested in learning about its veracity

Steps

1. Find all atomic, decontextualized claims from text
2. Retrieve only checkworthy ones

# EVALUATION METRIC

- **Lexical** similarity metrics are not enough for evaluating semantic similarity

| Sentence 1 | Sentence 2 | ROUGE-1 | ROUGE-L | Edit distance |
|---|---|---|---|---|
| Mount Everest is the highest mountain in the world. | The world's tallest peak is Mount Everest. | 0.588 | 0.235 | 40 |
| *Mount Everest* is the highest mountain in the world. | *Kilimanjaro* is the highest mountain in the world. | **0.824** | **0.824** | **12** |

# Claim extraction –evaluation

- **Lexical** similarity metrics are not enough for evaluating semantic similarity
- **Semantic** are better but still have problems with some edge cases such as negation

| Sentence 1 | Sentence 2 | ROUGE-1 | ROUGE-L | Edit distance |
|---|---|---|---|---|
| Mount Everest is the highest mountain in the world. | The world's tallest peak is Mount Everest. | 0.588 | 0.235 | 40 |
| *Mount Everest* is the highest mountain in the world. | *Kilimanjaro* is the highest mountain in the world. | **0.824** | **0.824** | **12** |

| Sentence 1 | Sentence 2 | BertScore |
|---|---|---|
| Tonda nebol na Mount Evereste. | Tonda bol na Mount Evereste. | 0.9 |

- **Lexical** similarity metrics are not enough for evaluating semantic similarity
- **Semantic** are better but still have problems with some edge cases such as negation
- **LLM as judge** might help mitigate these problems

| Sentence 1 | Sentence 2 | ROUGE-1 | ROUGE-L | Edit distance |
|---|---|---|---|---|
| Mount Everest is the highest mountain in the world. | The world's tallest peak is Mount Everest. | 0.588 | 0.235 | 40 |
| *Mount Everest* is the highest mountain in the world. | *Kilimanjaro* is the highest mountain in the world. | **0.824** | **0.824** | **12** |

| Sentence 1 | Sentence 2 | BertScore |
|---|---|---|
| Tonda nebol na Mount Evereste. | Tonda bol na Mount Evereste. | 0.9 |

# Claim extraction

Passat je nudný, ale celkem elegantní. Tohle je nudné, ale... No, prostě nejlepší, co se dá o na první pohled ID.7 říct, je, že sice vypadá jako čínská kopie ne moc hezkého auta, ale na jízdu do další rauchpauzy už se nenabíjí čtyři hodiny a má spoustu asistentů (které budete chtít vypnout, ale ono to nepůjde). Holt už teď asi budou všechna auta taková...

Passat je nudný, ale celkem elegantní. Tohle je nudné, ale... No, prostě nejlepší, co se dá o na první pohled ID.7 říct, je, že sice vypadá jako čínská kopie ne moc hezkého auta, ale na jízdu do další rauchpauzy už se nenabíjí čtyři hodiny a má spoustu asistentů (které budete chtít vypnout, ale ono to nepůjde). Holt už teď asi budou všechna auta taková...

Passat je nudný, ale celkem elegantní. Tohle je nudné, ale... No, prostě nejlepší, co se dá o na první pohled ID.7 říct, je, že sice vypadá jako čínská kopie ne moc hezkého auta, ale na jízdu do další rauchpauzy už se nenabíjí čtyři hodiny a má spoustu asistentů (které budete chtít vypnout, ale ono to nepůjde). Holt už teď asi budou všechna auta taková...

**Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny.**

**Volkswagen ID.7 je vybaven několika asistenčními systémy.**

**Asistenční systému u Volkswagen ID.7 se nedají vypnout.**

golden annotation

Passat je nudný, ale celkem elegantní. Tohle je nudné, ale... No, prostě nejlepší, co se dá o na první pohled ID.7 říct, je, že sice vypadá jako čínská kopie ne moc hezkého auta, ale na jízdu do další rauchpauzy už se nenabíjí čtyři hodiny a má spoustu asistentů (které budete chtít vypnout, ale ono to nepůjde). Holt už teď asi budou všechna auta taková...

**Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny.**

**Volkswagen ID.7 je vybaven několika asistenčními systémy.**

**Asistenční systému u Volkswagen ID.7 se nedají vypnout.**

golden annotation

**Volkswagen ID.7 se na krátkou jízdu nabíjí méně než čtyři hodiny.**

**Volkswagen ID.7 má spoustu asistentů, které se nedají vypnout.**

evaluated system

# Claim extraction

| | Volkswagen ID.7 se na krátkou jízdu nabíjí méně než čtyři hodiny. | Volkswagen ID.7 má spoustu asistentů, které se nedají vypnout. | [] |
|---|---|---|---|
| Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny. | 84 | 26 | 0 |
| Volkswagen ID.7 je vybaven několika asistenčními systémy. | 22 | 58 | 0 |
| Asistenční systému u Volkswagen ID.7 se nedají vypnout. | 29 | 75 | 0 |

# Claim extraction

| | Volkswagen ID.7 se na krátkou jízdu nabíjí méně než čtyři hodiny. | Volkswagen ID.7 má spoustu asistentů, které se nedají vypnout. | [] |
|---|---|---|---|
| Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny. | 84 | 26 | 0 |
| Volkswagen ID.7 je vybaven několika asistenčními systémy. | 22 | 58 | 0 |
| Asistenční systému u Volkswagen ID.7 se nedají vypnout. | 29 | 75 | 0 |

- Edit similarity
- BertScore
- LLM as judge

# Claim extraction

| | Volkswagen ID.7 se na krátkou jízdu nabíjí méně než čtyři hodiny. | Volkswagen ID.7 má spoustu asistentů, které se nedají vypnout. | [] |
|---|---|---|---|
| Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny. | 84 | 26 | 0 |
| Volkswagen ID.7 je vybaven několika asistenčními systémy. | 22 | 58 | 0 |
| Asistenční systému u Volkswagen ID.7 se nedají vypnout. | 29 | 75 | 0 |

- Edit similarity
- BertScore
- LLM as judge

# Claim extraction


**Volkswagen ID.7 se na krátkou jízdu nemusí nabíjet čtyři hodiny.**

**Volkswagen ID.7 je vybaven několika asistenčními systémy.**

**Asistenční systému u Volkswagen ID.7 se nedají vypnout.**

golden annotation

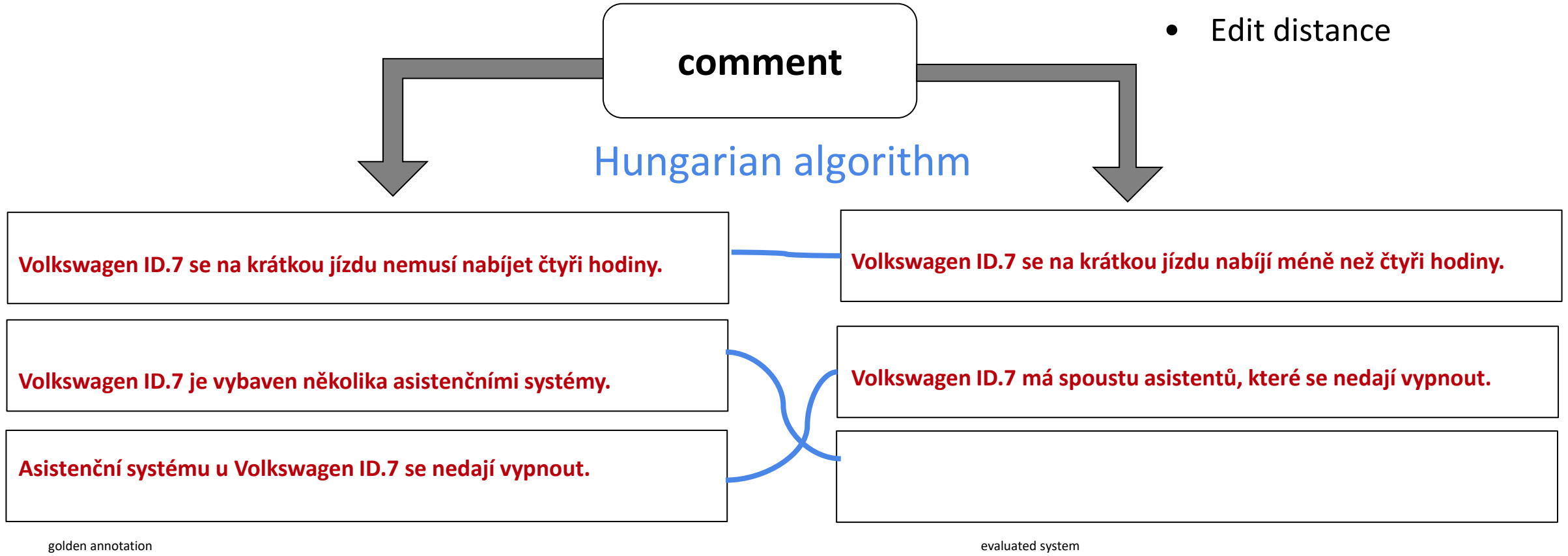
**Volkswagen ID.7 se na krátkou jízdu nabíjí méně než čtyři hodiny.**

**Volkswagen ID.7 má spoustu asistentů, které se nedají vypnout.**

evaluated system

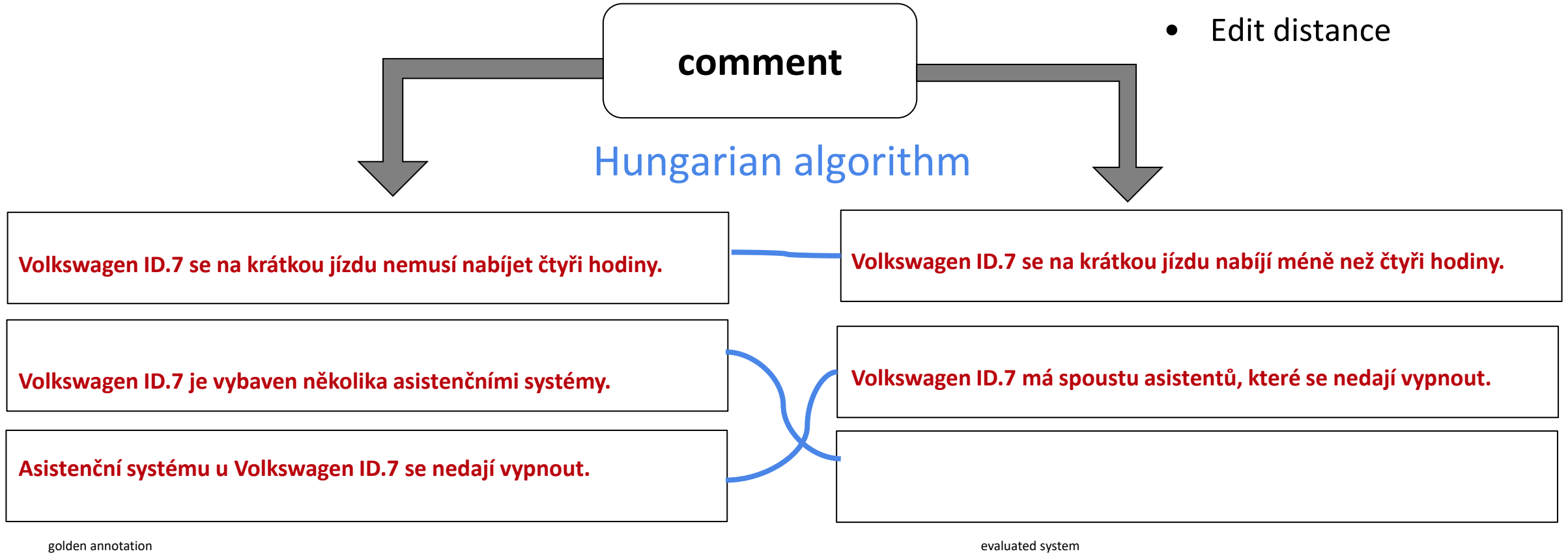
# Claim extraction

- LLM as judge
  - likert scale
- BertScore
- Edit distance



$$\textbf{Final Score} = \frac{1}{n}\sum_{i=1}^{n}(\text{Pair Score}_i)$$

- LLM as judge
  - likert scale
- BertScore
- Edit distance

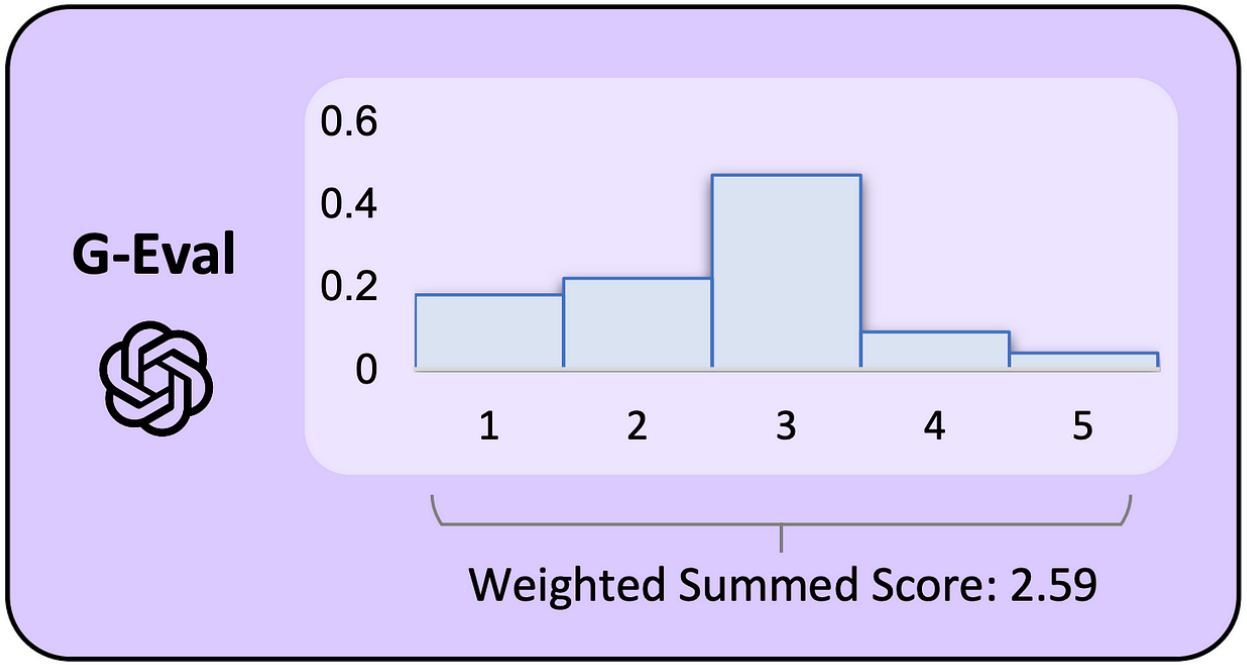


**Models**: phi 4, qwen 2.5

Logit based evaluation –fine-grained scores

**Models**: phi 4, qwen 2.5

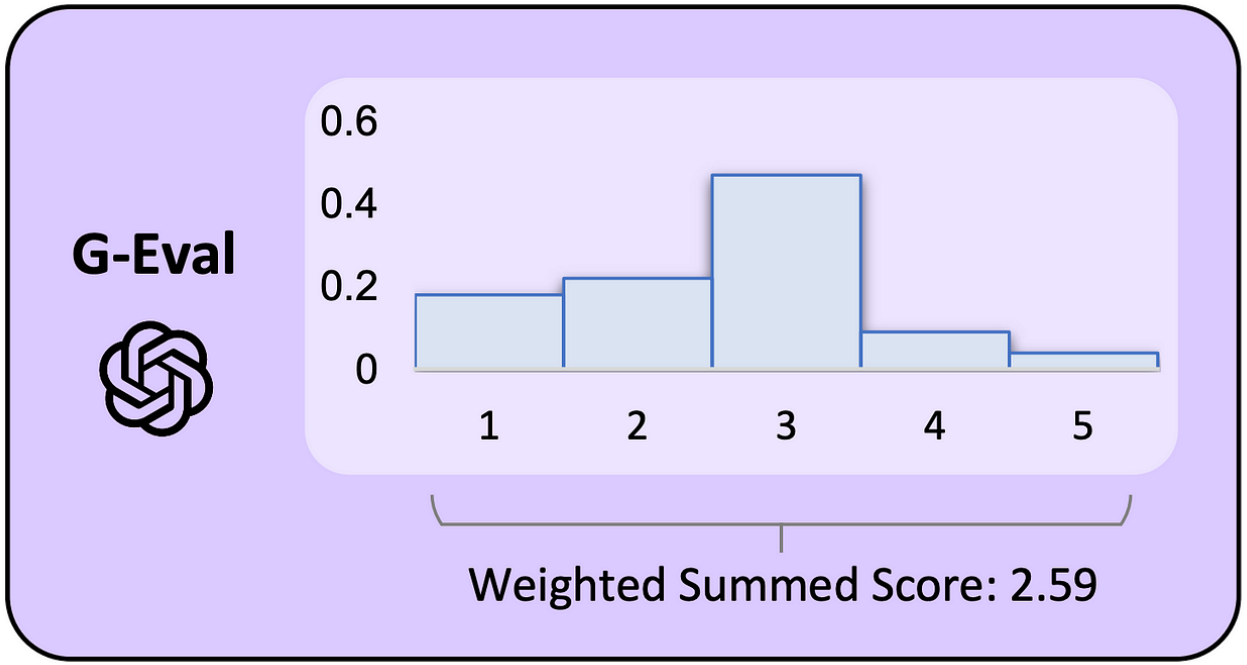
Logit based evaluation –fine-grained scores




G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment (Liu et al., EMNLP 2023)

**Models**: phi 4, qwen 2.5

Logit based evaluation –fine-grained scores




**Prompt:**

```
Tvým úkolem je porovnat dvojici tvrzení a vyhodnotit jejich
sémantickou podobnost a nakolik jsou fakticky v souladu.

Škála hodnocení:
0 - Úplně rozdílná tvrzení nebo jedno tvrzení chybí.
1 - Stejné téma, ale jiný nebo protikladný fakt.
2 - Tvrzení mluví o stejné veci, ale jedno z nich
    přidává nebo mění fakt, který zásadně ovlivňuje význam.
3 - Tvrzení mají téměř stejný význam, liší se jen
    drobným detailem který neovlivňuje zásadně význam.
4 - Totožná tvrzení nebo parafráze se stejným významem.


Příklad vstupu:
["Mary je pětiletá dívka.","Anna chodí do školky."]

Příklad výstupu:0

Výroky k porovnání:
```

G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment (Liu et al., EMNLP 2023)

Results of claim extraction evaluation (normalized to 0-100)

| Data | Model | BERTScore | LLM-P | LLM-Q | Edit Similarity |
|---|---|---|---|---|---|
| **CGCD** | Sonnet 3.7 | 52.8 | 41.3 | 40.9 | 35.6 |
| | **GPT 4o mini** | **60.2** | **45.3** | **43.8** | **37.9** |
| | GPT 4.1 | 54.3 | 43.4 | 42.2 | 36.7 |
| | GPT 4.1 (c) | 54.7 | 41.8 | 43.4 | 35.8 |
| **CSAD** | a1 vs a2 | 53.8 | 47.7 | – | 28.3 |
| | Sonnet 3.5 | 53.9 | 44.7 | – | 30.7 |
| | **GPT 4o mini** | **58.3** | **49** | – | **34.9** |
| | GPT 4.1 | 55.2 | 46 | – | 32.7 |
| | GPT 4.1 (c) | 52.6 | 44.7 | – | 31.2 |

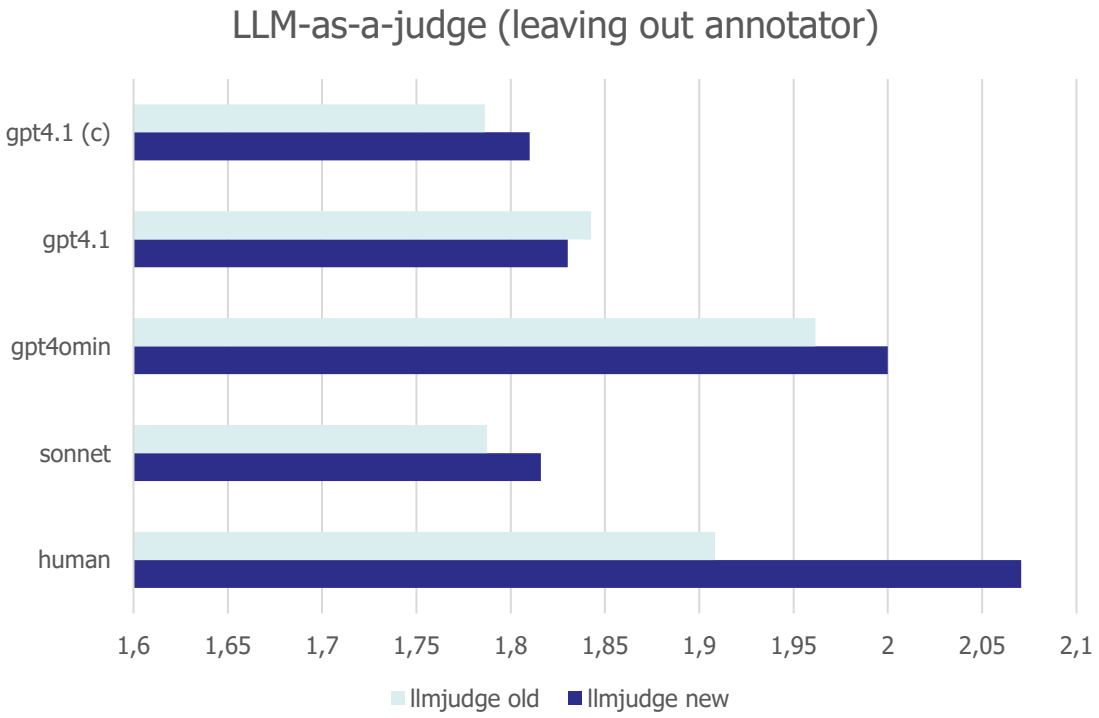
Table: Results of evaluation of 4 different models (GPT 4.1 (c) means that model was provided context in a way of source text url) using our proposed evaluation method with 4 different metrics (For LLM-P – phi4 and LLM-Q – Qwen 2.5 were used as models for llm as judge)

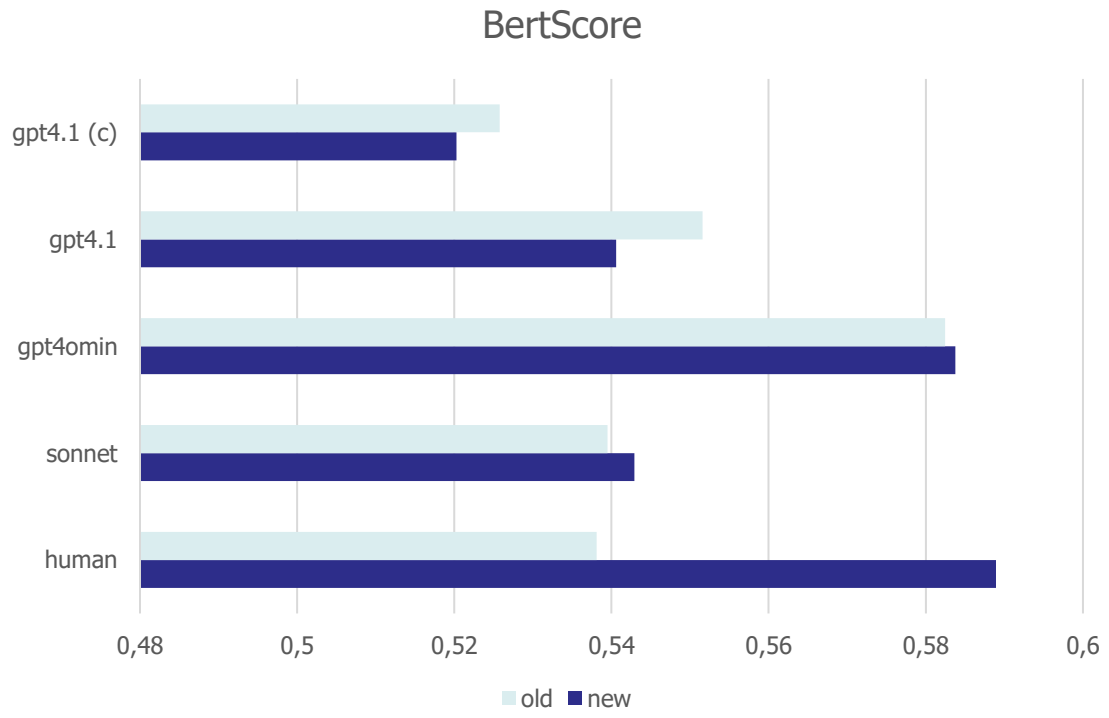
# Claim extraction – IAA

- leave-one-out annotator evaluation
- indentify outlier and inspect their annotations
- recalculate the score without their data

# Claim extraction – IAA

- leave-one-out annotator evaluation
- indentify outlier and inspect their annotations
- recalculate the score without their data



**Separate task**

# Claim extraction -subtask evaluation

**Separate task**

1. Right number of claims

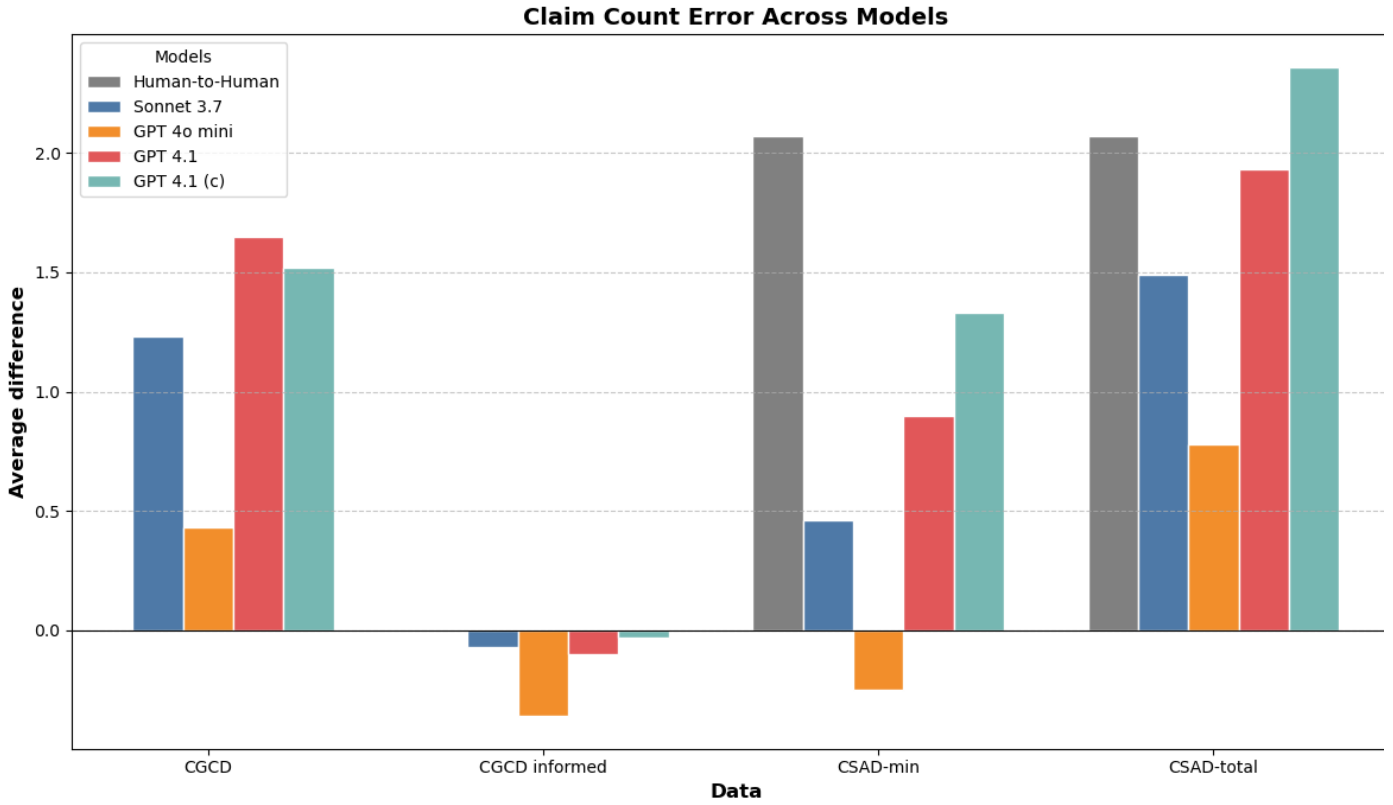



Table: Mean Difference in Number of Claims Between Model and Human Annotators (measuring subtask 1)

# Claim extraction -subtask evaluation

**Separate task**

1. Right number of claims
2. Extracting correct claims

Informed: model was told (in prompt) the expected number of checkworthy claims

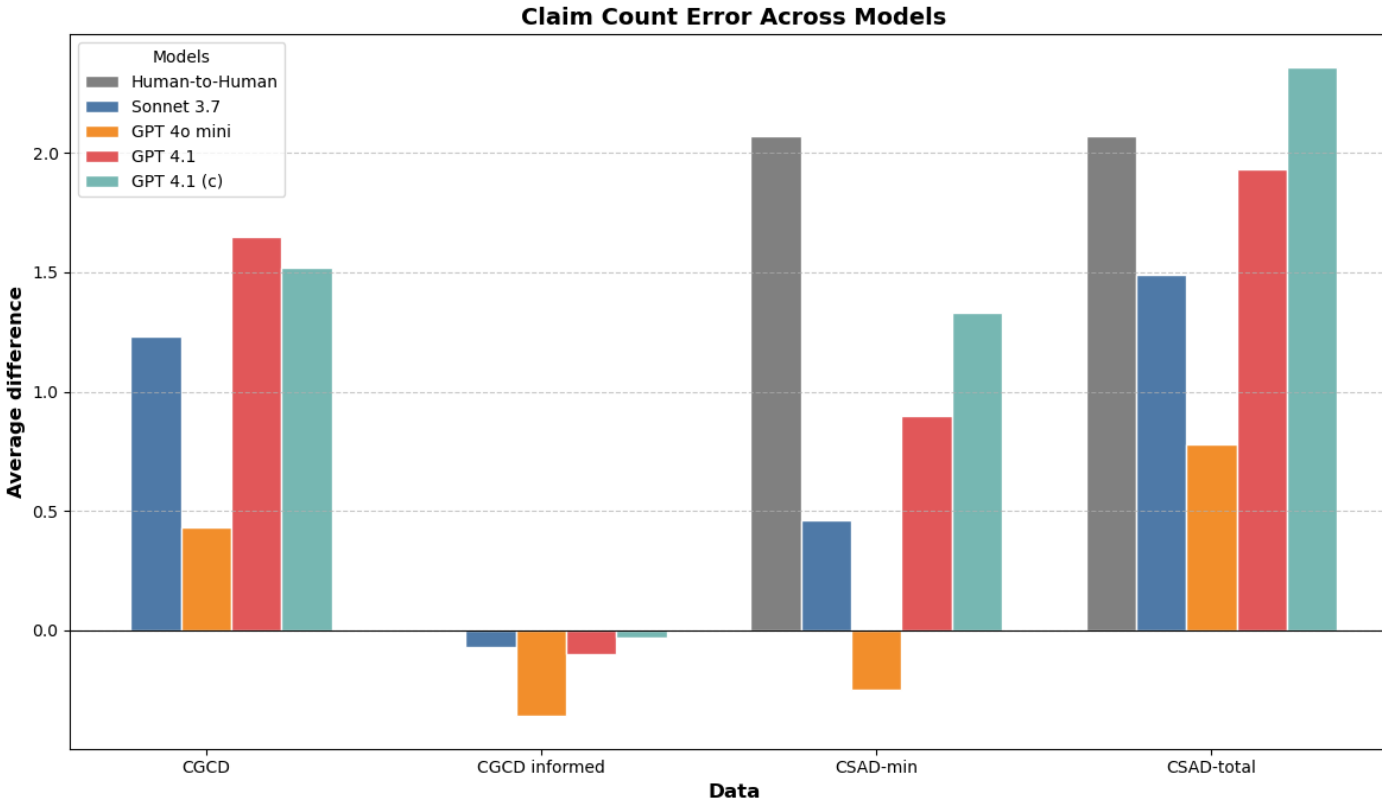
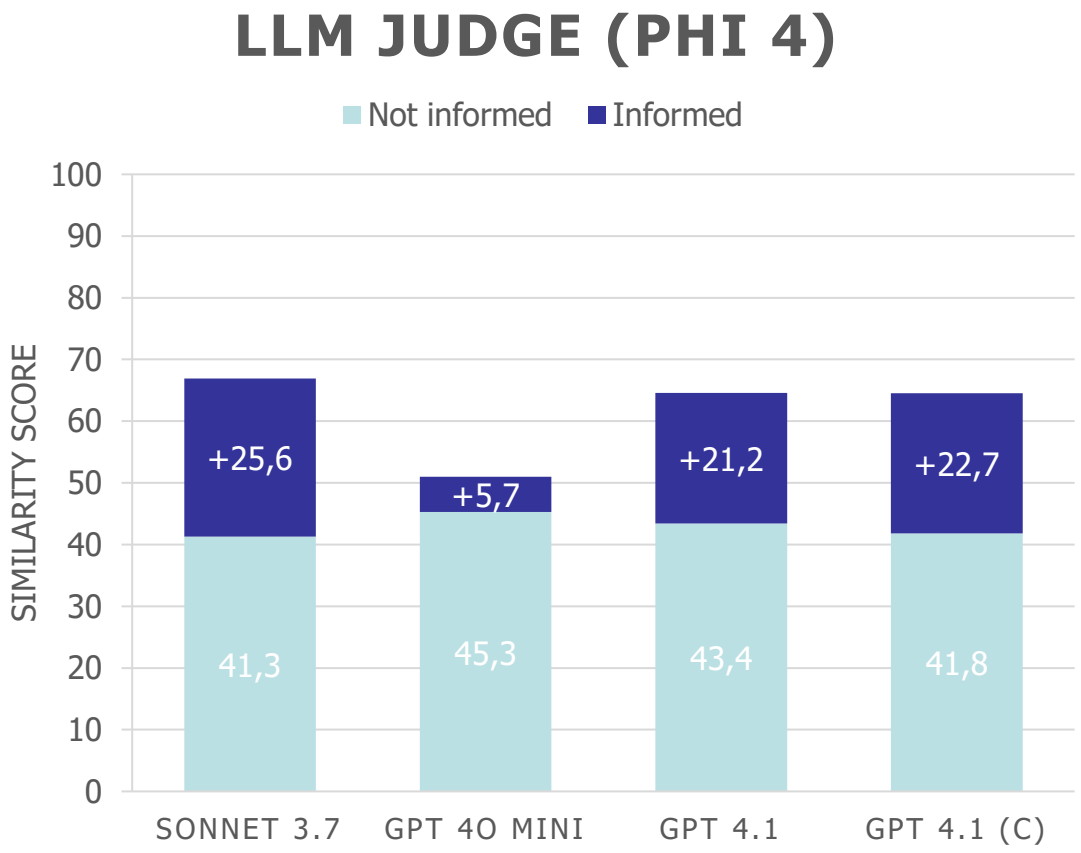


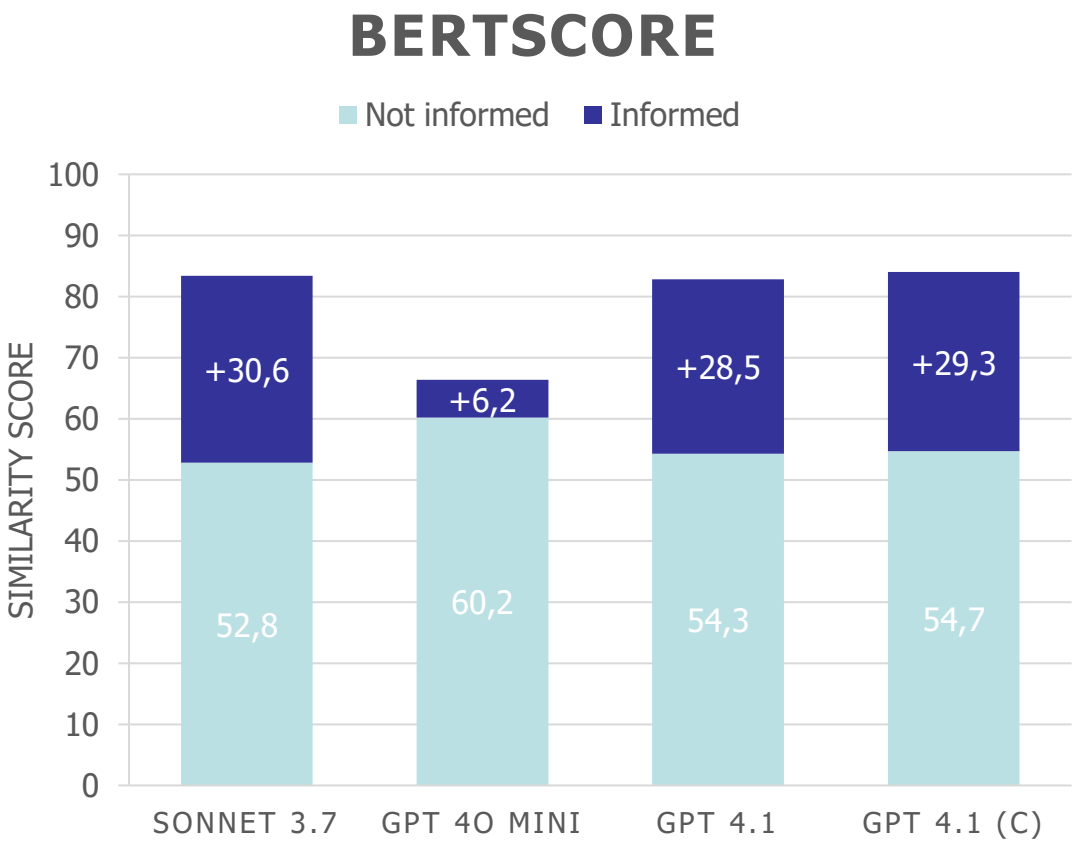

Table: Mean Difference in Number of Claims Between Model and Human Annotators (measuring subtask 1)

# Claim extraction - subtask evaluation

- significant improvement after informing about the expected number of claims
- penalization is too strict

- significant improvement after informing about the expected number of claims
- penalization is too strict

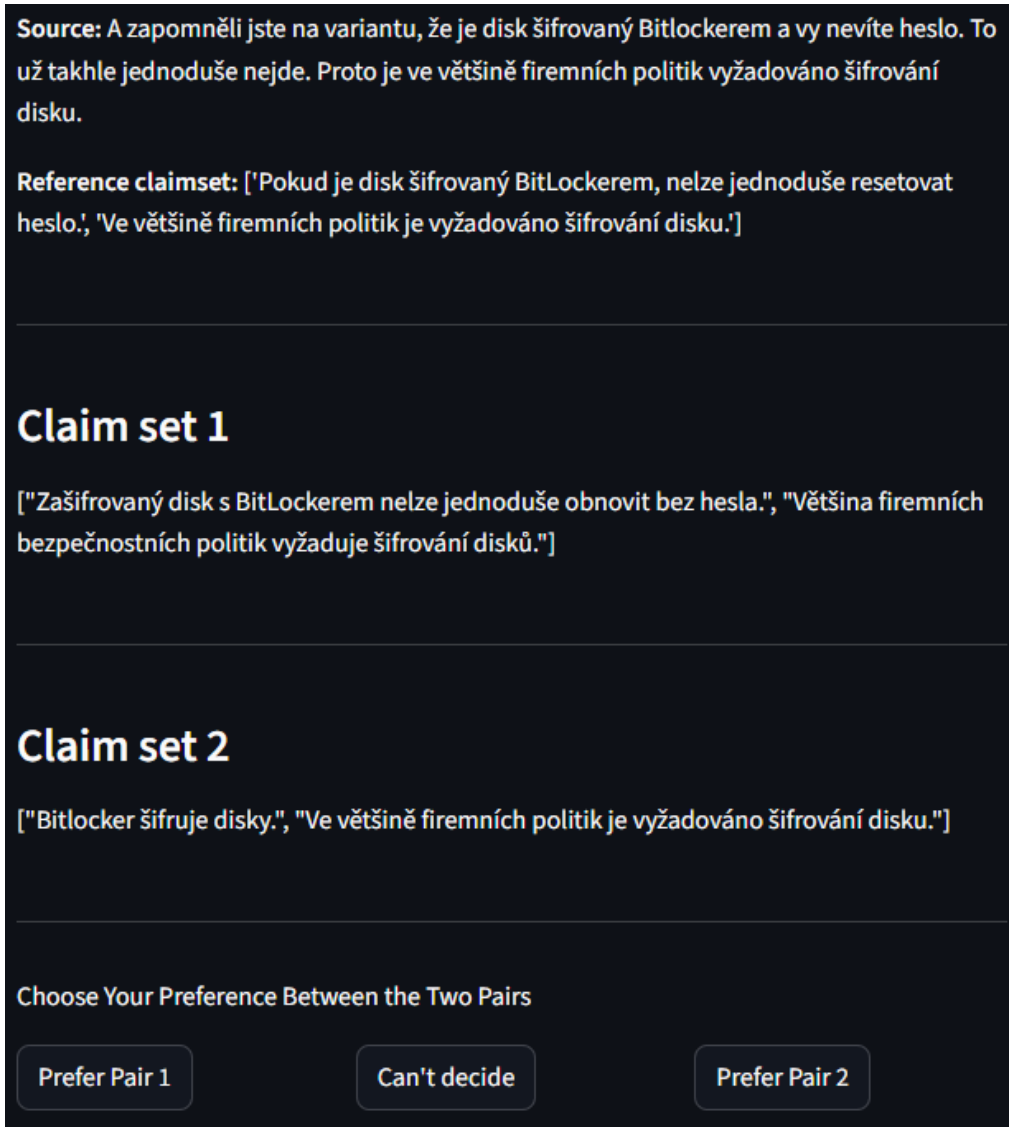


# Claim extraction -human evaluation

Human preference evaluation

- 120 preference samples for each group
- Voting options: 1, 2, can’t decide



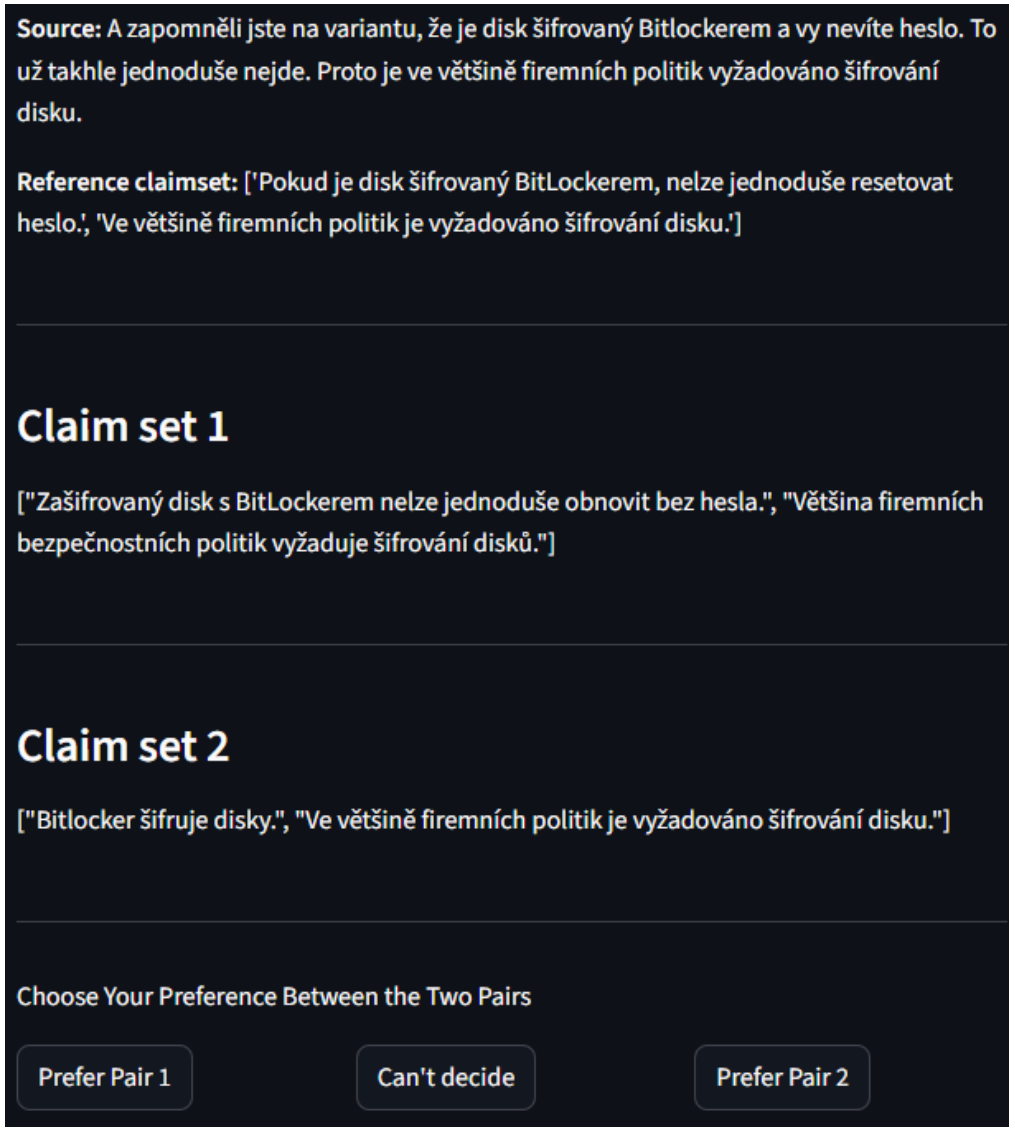
# Claim extraction -human evaluation

Human preference evaluation

- 120 preference samples for each group
- Voting options: 1, 2, can’t decide

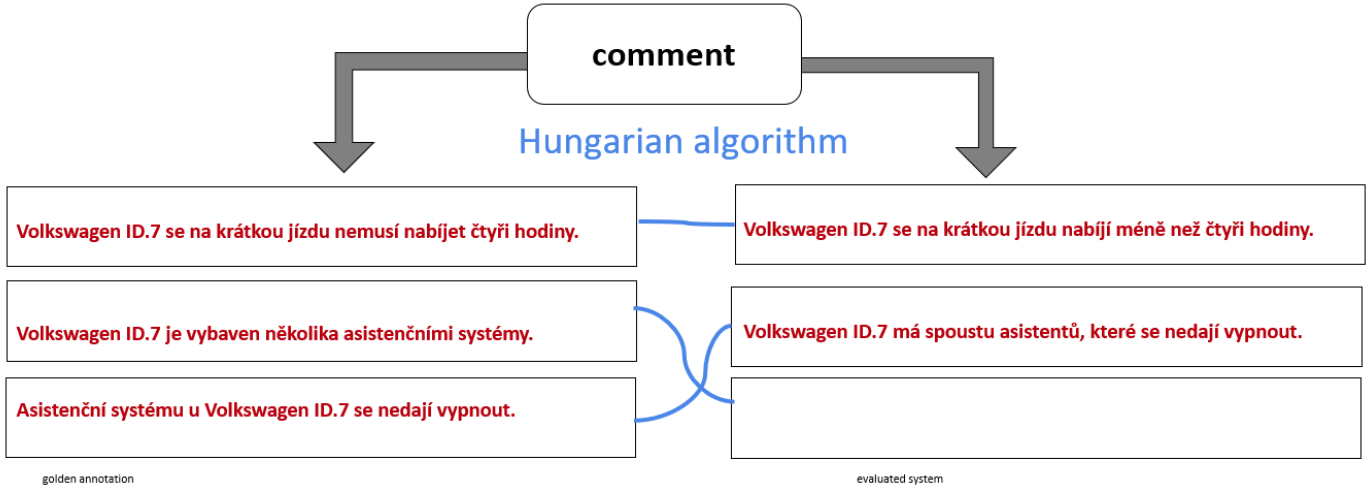
| Model | NINF | INF | Pair 1 | Pair 2 |
|---|---|---|---|---|
| Sonnet 3.7 | 40 | 53 | 50 | 52 |
| GPT 4o mini | 38 | 20 | - | - |
| GPT 4.1 | 28 | 51 | - | 41 |
| GPT 4.1 (c) | 40 | 41 | 39 | - |

Table: Number of preference points obtained by each model (NINF – non informed models group, INF – informed models group Pair1 – Sonnet 3.7 vs GPT 4.1 with context, Pair2 – Sonnet 3.7 vs GPT 4.1)



# Thank you for your attention!

## Proposed way to ealuate document-level claim extraction




## Evaluated metrics agreement with humans

| Model | NINF | INF | Pair 1 | Pair 2 |
|---|---|---|---|---|
| Sonnet 3.7 | 40 | 53 | 50 | 52 |
| GPT 4o mini | 38 | 20 | - | - |
| GPT 4.1 | 28 | 51 | - | 41 |
| GPT 4.1 (c) | 40 | 41 | 39 | - |

## Identified problematic part of claim extraction evaluation