LLMs as Span Annotators:
A Comparative Study of LLMs and Humans

Zdeněk Kasner[1], Vilém Zouhar[2], Patrícia Schmidtová[1], Ivan Kartáč[1], Kristýna Onderková[1], Ondřej Plátek[1], Dimitra Gkatzia[3], Saad Mahamood[4], Ondřej Dušek[1], Simone Balloccu[5]

[1] Charles University [2] ETH zürich [3] Edinburgh Napier University [4] trivago [5] Technische Universität Darmstadt

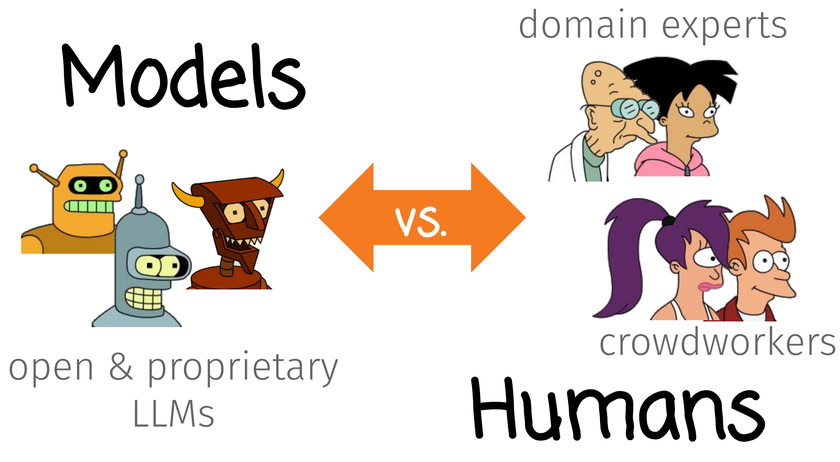
TYPO

LLMs can **anotate** text spans as well as skilled crowdworkers, but **human experts are better**.

OVERGENERALIZATION

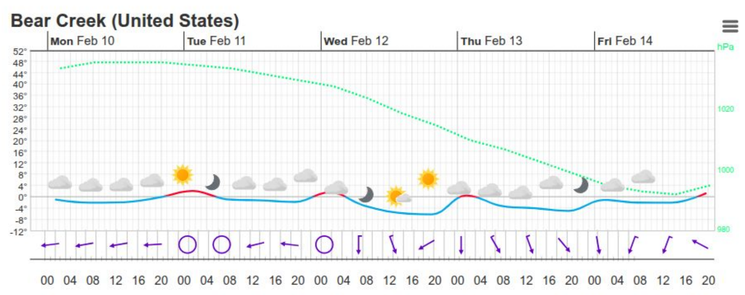
Task: Given a text & list of categories, **locate and categorize the spans** in the text.

Task	Description	Input X	Text Y with annotations A *(category, span, reason)*
D2T-Eval	***Detecting errors in data-to-text.*** *Custom annotation from Prolific.*	Mon Tue Wed	Skies will be C **mostly clear** *Rain on Mon & Wed*, but NC **winds will remain strong** *Wind speed data is missing.* .
MT-Eval	***Detecting errors in machine translation.*** *WMT 2024: English → 9 langs.*	Der schnelle braune Fuchs springt über den faulen Hund.	The quick brown fox MN **jump** *Third person singular* over the lazy MJ **fox** *'Hund' translates to 'dog'* .
Propaganda	***Detecting propaganda techniques.*** *Da San Martino et al. (2019): various online sources*	∅	AA **Study Finds** That Driving Car Is More Efficient than Biking *Appeal to a 'study'*



Research question: Can we reliably **automate span annotation with LLMs?**

Input: weather data (JSON for models, visual chart for humans) & generated forecast

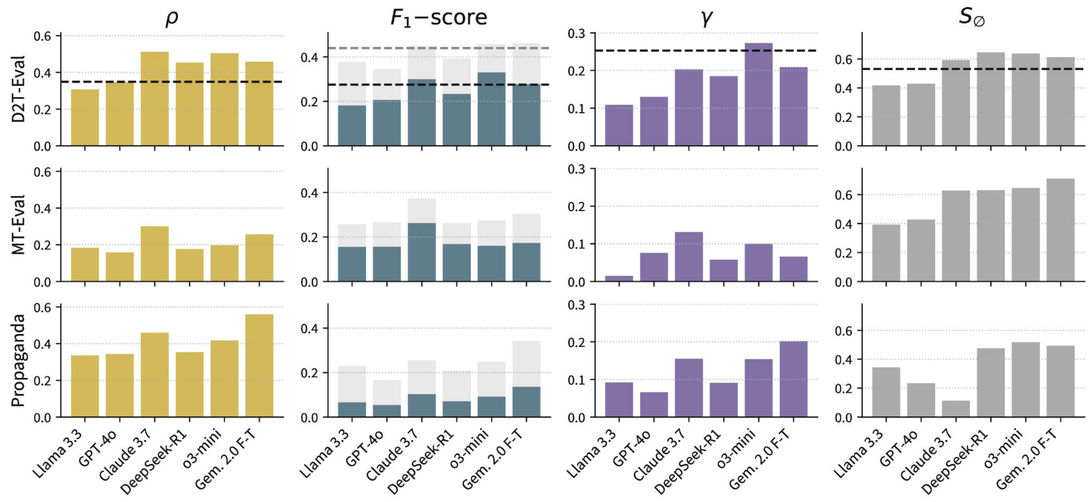


Example

human	Claude 3.7
Bear Creek will experience mostly overcast skies for the next five days, with a chance **of occasional showers on Wednesday and Thursday.** Temperatures will remain relatively cool, hovering around freezing point or slightly below. Winds will be light to moderate throughout the period, **with some gusts expected on Friday.** The overall weather pattern suggests a generally cloudy and **potentially wet week.**	Bear Creek will experience mostly overcast skies for the next five days, **with a chance of occasional showers on Wednesday and Thursday**. Temperatures will remain relatively cool, hovering around freezing point or slightly below. Winds will be light to moderate throughout the period, **with some gusts expected on Friday**. The overall weather pattern suggests a generally cloudy and **potentially wet week**.

Output: errors in the generated forecast (**Contradictory**, **Misleading**)

- **Pearson ρ** : span counts
- **γ score** (Mathet et al., 2015) } position & category
- **F1** (w/overlaps) }
- **$S_{ø}$-score**: ex. with no spans



How well are the spans **aligned**?

Evaluation

How **efficient** is each approach?

Are the spans **correct**?

Annotator	Cost	Time
	$/1k out	s/out
crowdworkers	500	129.1
Llama 3.3 70B	-	21.6
DeepSeek-R1 70B	-	227.5
Claude 3.7 Sonnet	10.5	9.8
o3-mini	**3.6**	21.8

Takeaways

Low-to-moderate agreement → LLMs **cannot** replace human annotators straight away.

However, depends on the **domain and annotator expertise**.

In some domains, LLMs **can replace skilled annotators.**

LLMs are also more **cost- and time-efficient** than human annotators.

Paper & data explorer available at `llm-span-annotators.github.io`

Presented at Multilingual and Multicultural Evaluation (MME) Workshop, EMNLP 2026, Rabat, Morocco

Funded by the European Union (ERC, NG-NLG, 101039303). Supported by the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and the Charles University Research Centre program No.~24/SSH/009. It used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).