# Large Language Models as Assessors: On the Impact of Relevance Scales

Riccardo Zamolo,[1] Riccardo Lunardi,[1] Michael Soprano,[1] Gianluca Demartini,[2] Stefano Mizzaro,[1] and Kevin Roitero[1]
[1]University of Udine [2]University of Queensland

**This study explores how different scales and their conversions affect relevance assessments provided by LLMs across multiple prompting strategies and model sizes.**

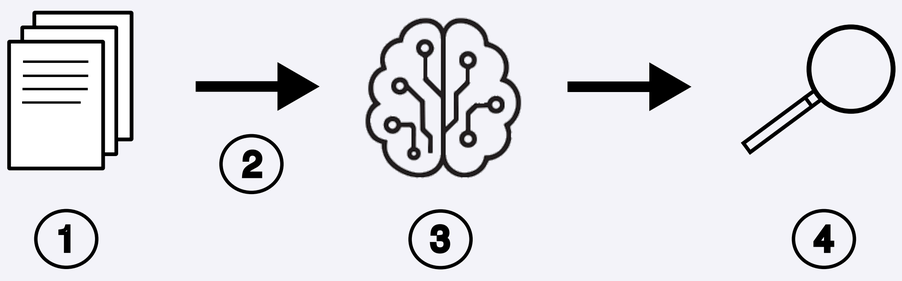
## Motivations

- Evaluate LLMs as alternative to **costly** human relevance judgments in IR
- Analyze **how** scales **impact** the accuracy and stability of LLM-based assessments
- Examine **if** scale conversion **affects** LLM reliability and alignment with human labels

## Research questions

**RQ1** What is the **effect**, if any, of **changing** the scale of relevance judgment expression? What is the **impact** of LLM-based relevance judgments on system ranking?
**RQ2** What **differences** arise between using a **target** scale directly and applying **conversions** from another scale?

## Experimental Pipeline




**1 Dataset** 18 topics and 3,550 documents after pre-processing, resulting in **3,881 topic-document pairs**, taken from a subset of TREC-8. Judgments are both collected from crowd workers and NIST expert assessors.

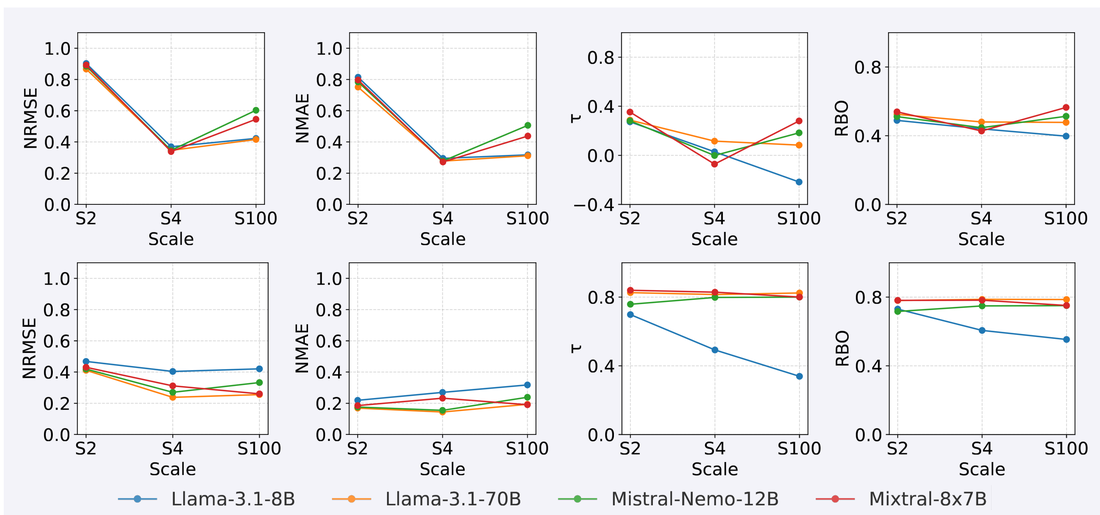
**2 Prompts** A total of **12 prompts** are used:
- 4 taken from previous work and identified by the name of the main author
- 8 original prompts, labeled according to the prompt engineering techniques applied

**Scales** **4 evaluation scales**:
- Binary ($S_2$), which corresponds to NIST experts judgments
- 4-level ordinal ($S_4$), to ensure comparability with prior work
- 8-level ordinal ($S_8$), to test any potential effects of data contamination in LLMs
- Fine-grained numerical scale ($S_{100}$), to ensure comparability with prior work

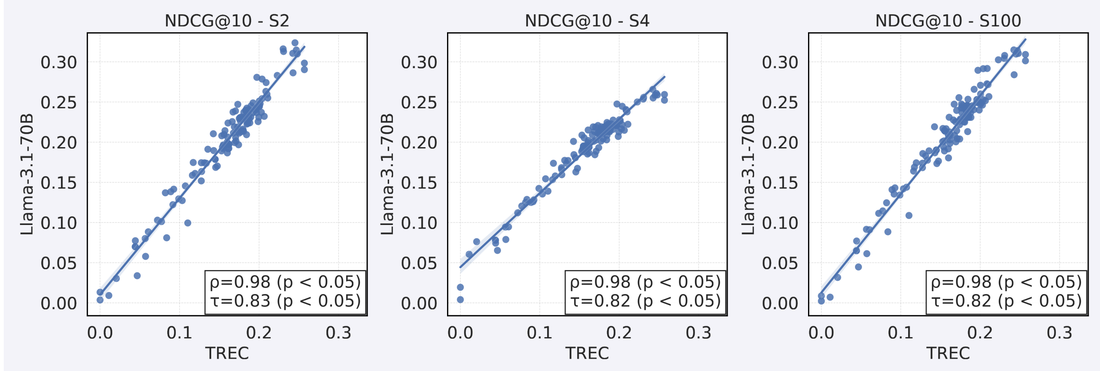
**3 Models** **4 state-of-the-art LLMs** employed:
- Llama-3.1-8B
- Llama-3.1-70B
- Mistral-Nemo-12B
- Mixtral-8x7B

**4 Results** on Llama-3.1-70B, which is the **most** stable model:

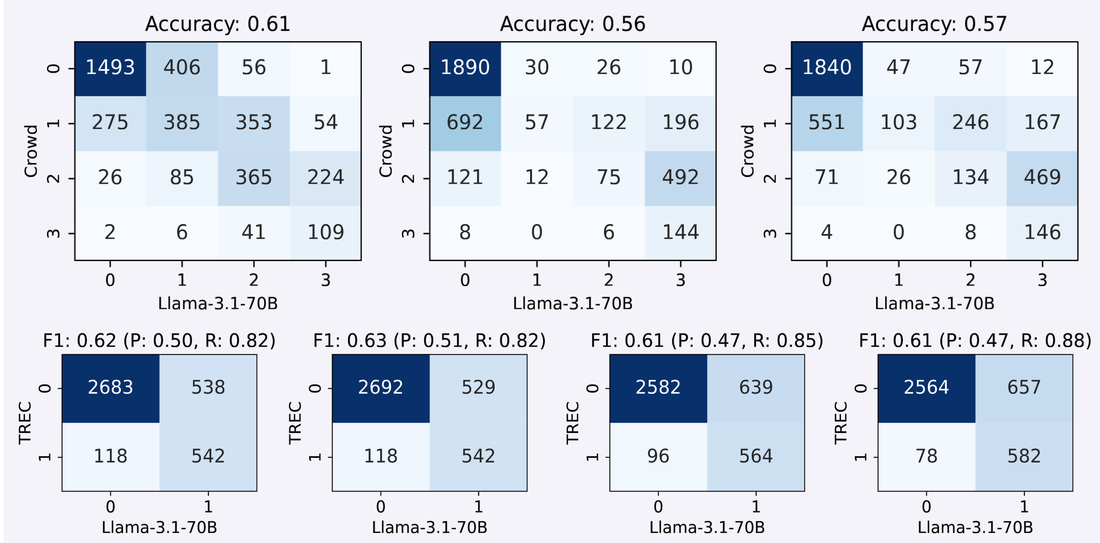
**RQ1** - $S_4$ provides the **best** balance of accuracy and ranking stability, particularly when paired with more effective **structured** prompting, which mitigate scale-related effects.




LLM-based judgments show a **strong** linear correlation and ranking agreement with TREC scores across **all** scales. $S_4$ provides the **most** stable and consistent system rankings, exhibiting the **highest** level of alignment.




**RQ2** - Native scales are **more** accurate and align **better** with human judgments than converted ones. While some data contamination is present, its impact on the results remains **minimal**.




## Limitations

- No expert ground truth available for **all** scales
- **One-dimensional** notion of relevance
- Full focus on **pointwise** assessments

## Future Work

- Extend to **more** datasets, prompts, scales and models
- Consider **multidimensional** notion of relevance
- Explore **alternative** assessment formats

## Resources

- Repository at `https://osf.io/xkvm6/`
- Source code, results, prompts, and scale definitions

EGIR 2026