



Why Large Language Models can Secretly Outperform Embedding Similarity in Information Retrieval

Matei Benescu Ivo Pascal de Jong

Department of Artificial Intelligence, Bernoulli Institute – University of Groningen

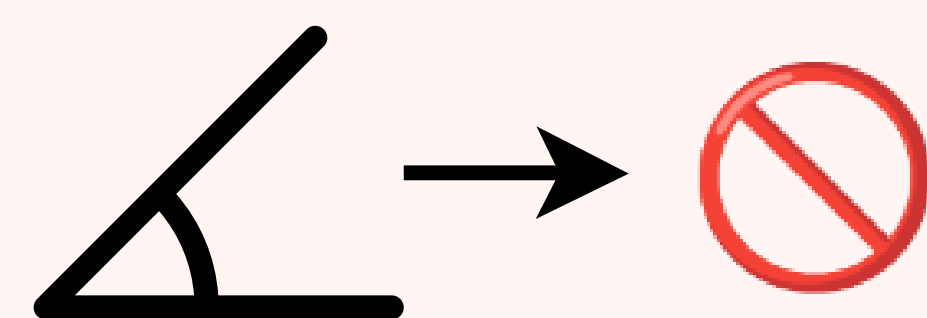
Problem: Short-Sighted Similarity

Information Retrieval finds documents that are relevant to a query. Modern systems use **embedding similarity** between vector embeddings as a proxy for relevance. Similarity can be a “**short-sighted**” proxy for relevance causing specific failure modes.

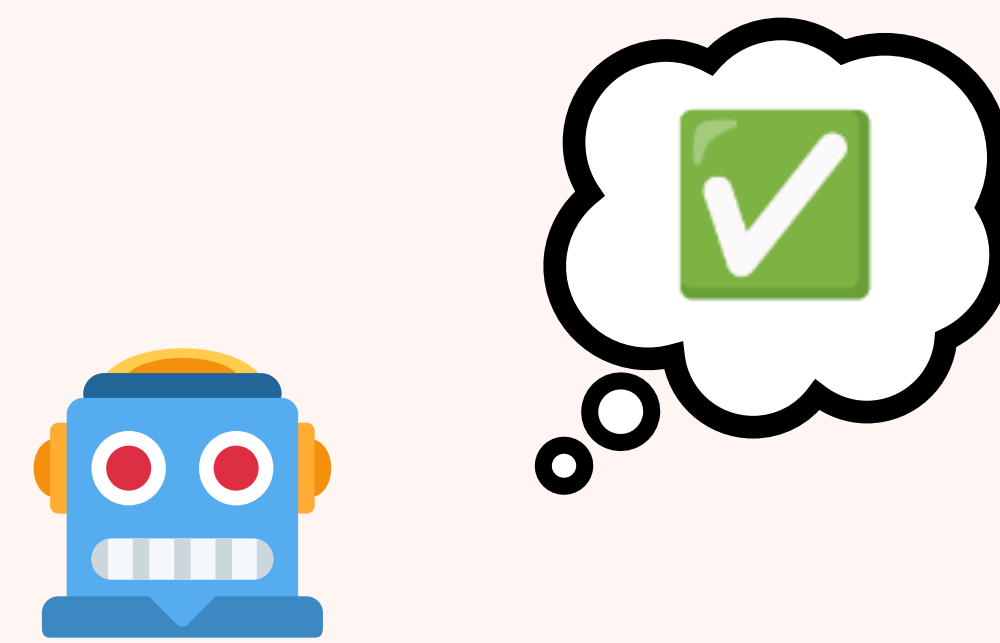
Do these match?

query:
difference between
mcdouble and double
cheeseburger

document:
one slice of cheese
instead of two. It's
twenty cents more.



Embedding
Similarity



Relevance
Reasoning

Figure 1. **Short-sightedness problem.** Embeddings (NERS) only look at similarity as a proxy for relevance. An LLM (LLM-RJS) can solve this by reasoning about relevance.

Main Takeaways

- Evaluated on **human annotation**, NERS performs best at **minimal inference cost** (Fig. 3).
- But LLM-RJS is able to find relevant documents that the humans missed (Fig. 4). Humans and NERS are both short-sighted (Tab. 1).
- Embedding similarity is used in many domains, but is intrinsically short-sighted. Modern models with **reasoning can overcome this fundamental short-sightedness problem**.

Experimental Setup

We compare state-of-the-art Neural Embedding Retrieval Systems (NERS) against LLM-based Relevance Judgement Systems (LLM-RJS) with and without reasoning. We test several proprietary and open-source models on TREC-DL 2019 passage retrieval.

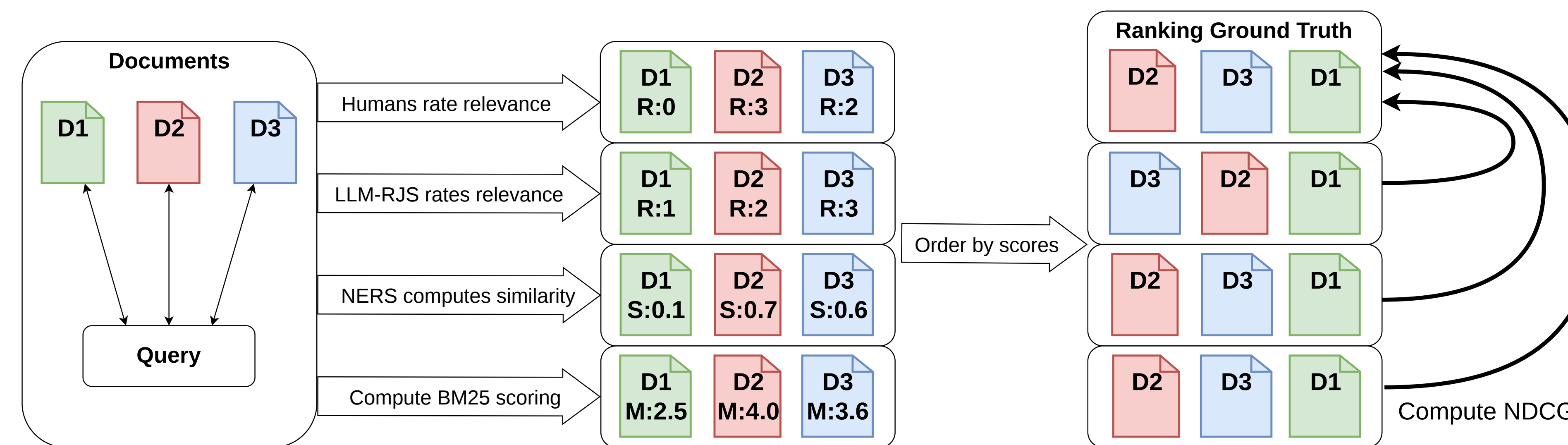


Figure 2. Documents are ranked based on the relevance score with respect to the query. NDCG measures **similarity to human ground-truths**.

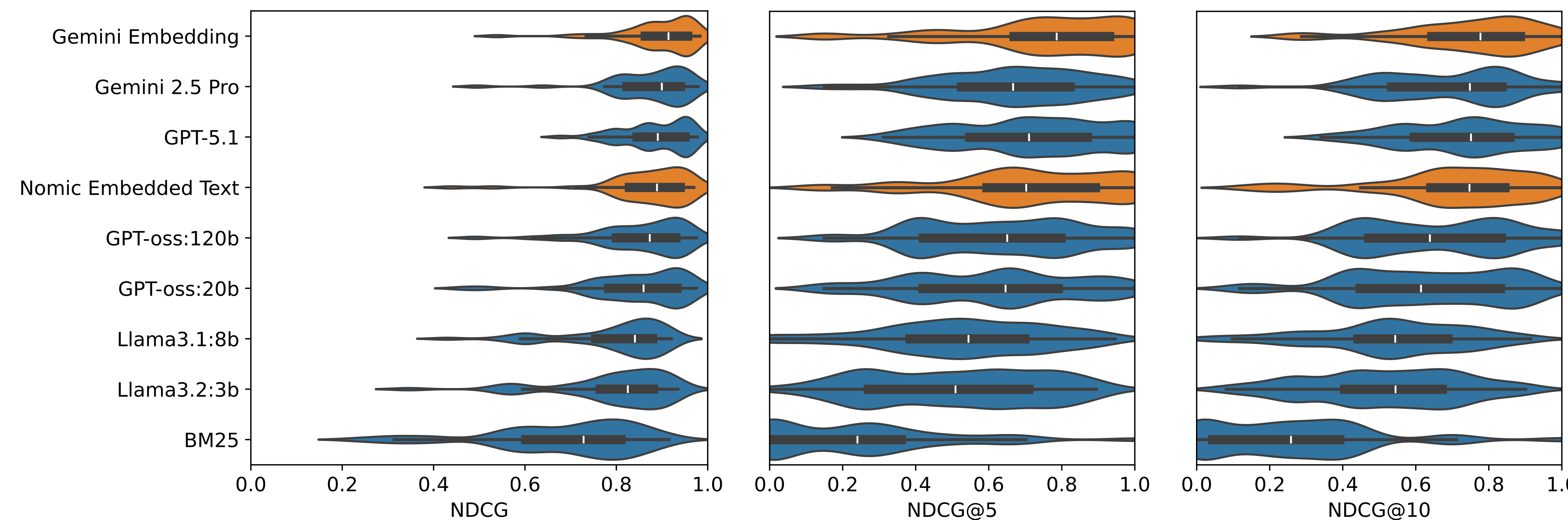


Figure 3. Short-sighted NERS models match marginally better with **human annotated** ground-truths than comparable LLM-RJS models.

Query	Passage	Human	LLM-RJS	NERS
do goldfish grow	Average ones are four to six inches (10–15 cm) long, but they can grow up to 7 inches [...]	0	3	0.35
example of monotonic function	[...] Since then, this proportion has grown steadily and monotonically with each successive election to 43% in 1996, 54% in 2000, 58% in 2004 [...]	0	3	0.14
how long is life cycle of flea	[...] Adults live about 7 to 10 days and deposit eggs on the moist soil surface or in soil cracks [...]	0	3	0.38

Table 1. Examples where LLM-RJS sees maximal relevance but **human annotations** do not. Of the 94 such cases, only 5 were genuine false positives.

The Secret

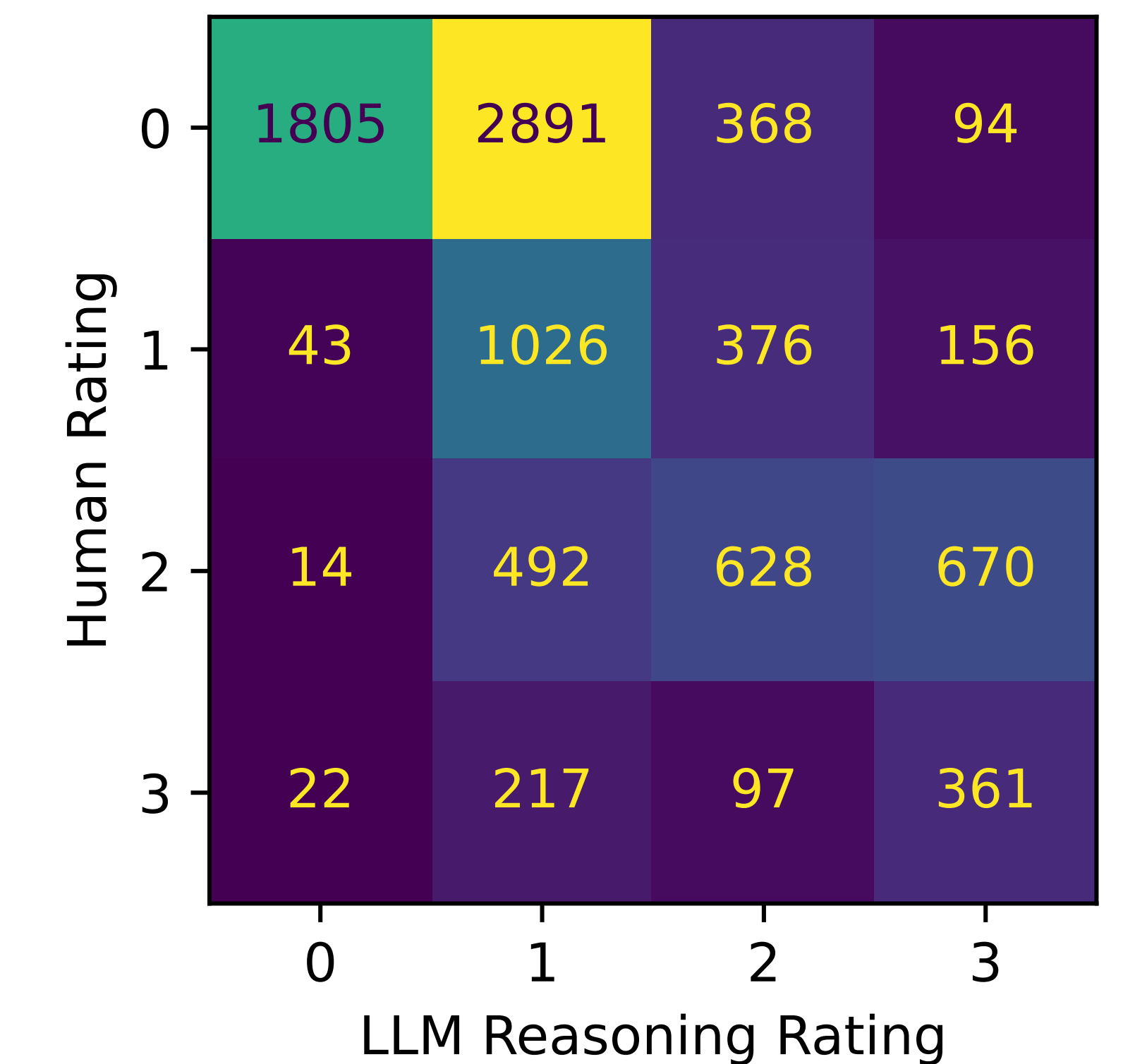
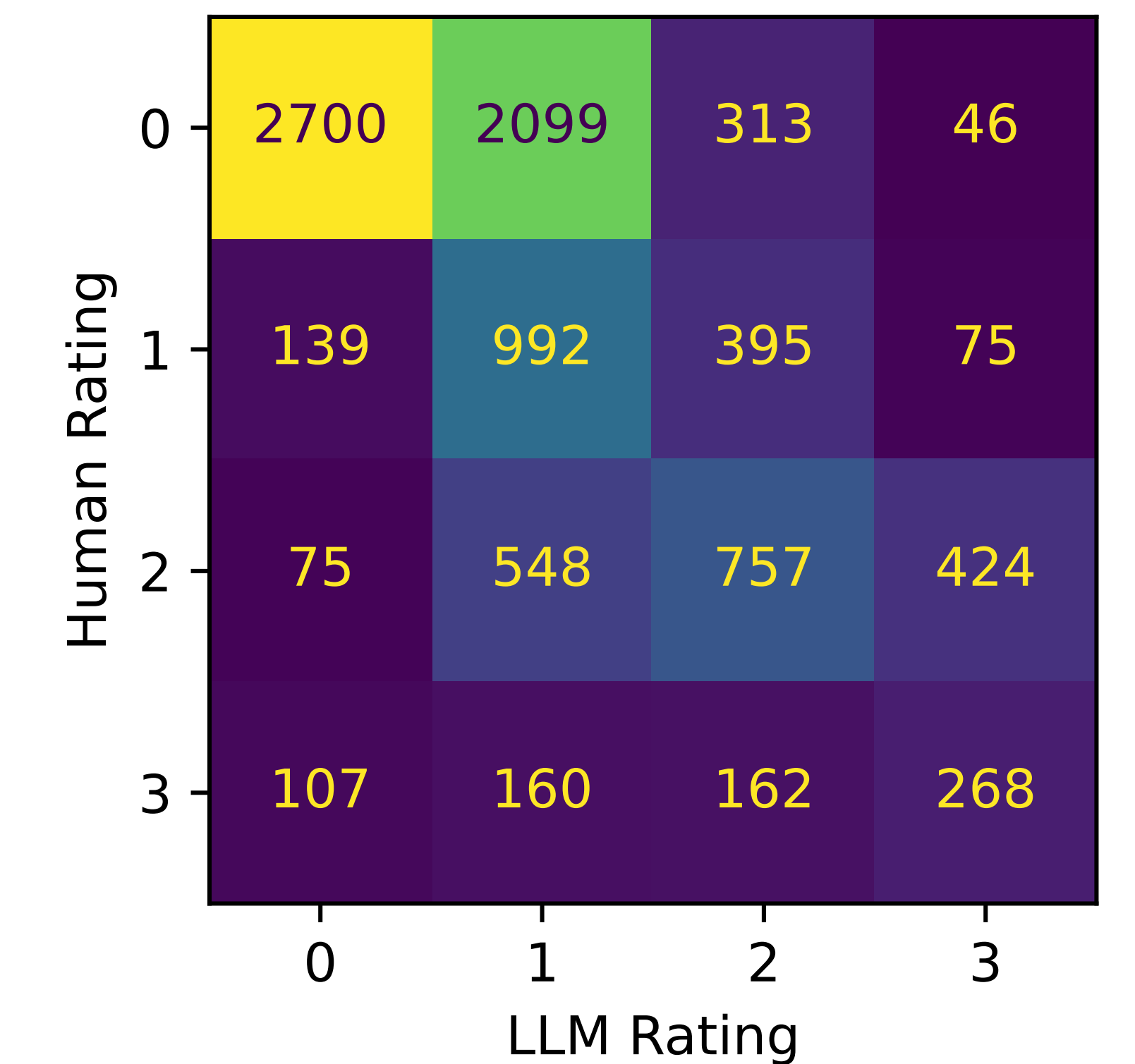


Figure 4. LLM-RJS are more likely to see relevance where humans do not. This is because **humans are also short-sighted!**