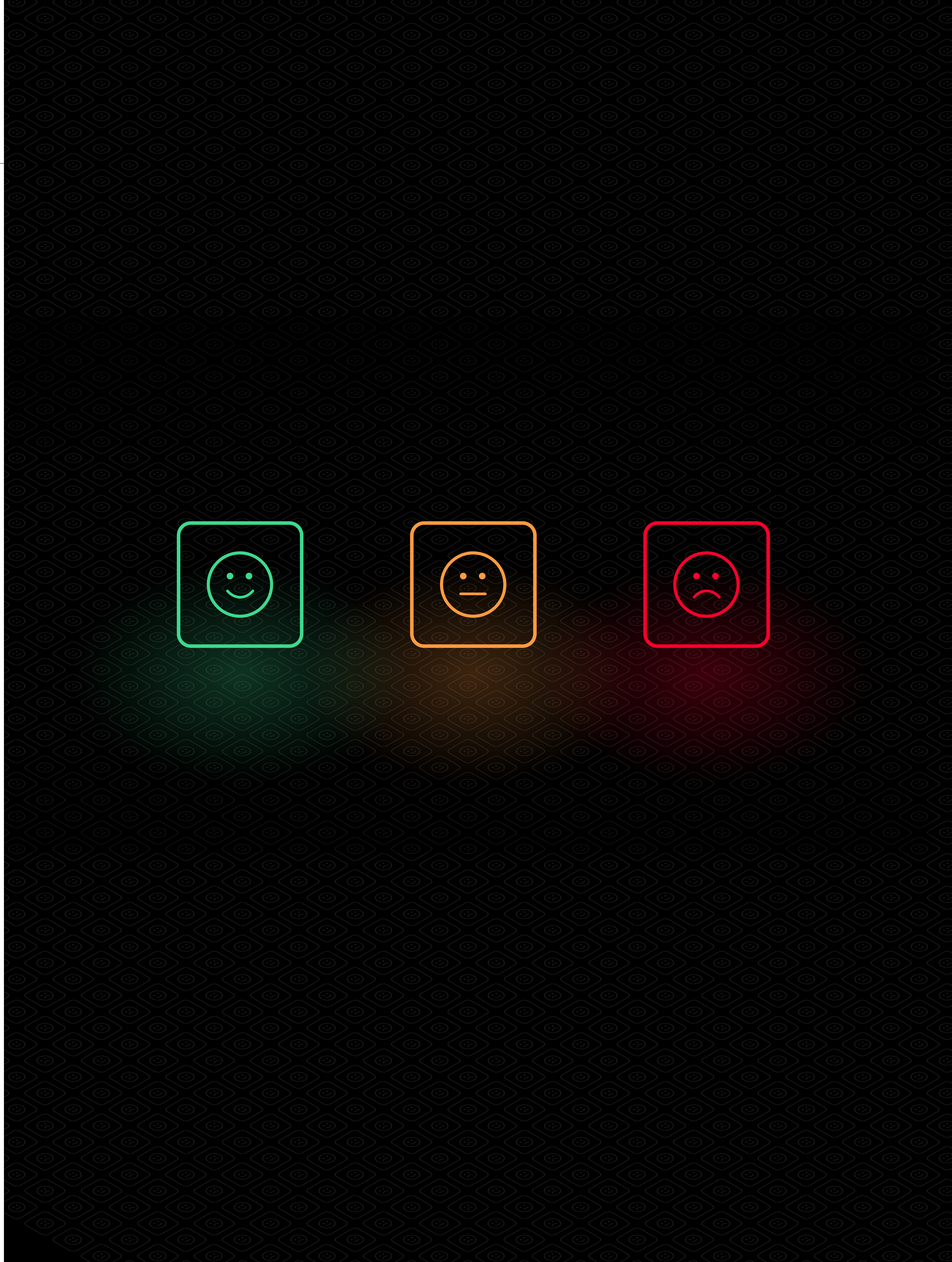
EVALUATING COST-ACCURACY TRADE-OFFS IN MULTIMODAL SEARCH RELEVANCE JUDGEMENTS

Silvia Terragni, Cuong Hoang, Joachim Daiber, Pallavi Gudipati, Pablo N. Mendes

Agenda

1. Introduction
2. Related Work
3. Methodology
4. Results
5. Conclusion



Introduction - Why is search relevance evaluation hard?

A judge needs to:

1. Understand the intent behind the search query
2. Interpret product relevance based on multiple features including the title, description, images, category, color, price
3. Take the use-case into consideration - eg. E-commerce vs News articles
4. Interpret and integrate information from various attributes across different modalities
5. Handle low quality data or data with missing attributes

Introduction - Why is search relevance evaluation hard?

1. Human annotation is reliable but costly and time-consuming
2. Large Language Models (LLMs) and Multimodal Language Models (MMLMs) are a viable alternative for producing relevance judgements
3. LLMs-as-judges can unlock higher relevance judgement throughput at a fraction of the cost

We evaluate the following:

1. Is LLM performance use-case dependent?
2. Is there a clear winner?
3. Is multimodal support necessary for search relevance judgement in multimodal search?
4. What models offer the optimal cost-accuracy trade-offs?

Related Work

1. Prometheus - a 13-billion parameter LLM designed to evaluate long texts using customized scoring rubrics provided by users
2. JudgeLM - fine-tuned LLMs as scalable judges to evaluate other LLMs effectively in open-ended tasks
3. Chen et al. assess judges through a new benchmark - performance in tasks such as Scoring Evaluation, Pair Comparison and Batch Ranking
4. Yang et al. investigates the relevance estimation of Vision-Language Models (VLMs), including CLIP, LLaVA, and GPT-4V

Methodology

Our pipeline consists of three stages:

1. *Data Collection:* Search results from 3 datasets
2. *Human Annotation:* 2 trained human annotators assigned relevance grades
3. *Model Evaluation:* Range of LLMs and MLLMs to generate relevance judgements

Methodology - Datasets

3 datasets - *Fashion, Hotel Supplies,* and *Design*

1. Fashion - H&M Personalized Fashion Recommendations (publicly available)
2. Hotel Supplies - E-commerce search for hotel supply products (proprietary)
3. Design - Social media search for design assets (proprietary)

Dataset	Total Number of Search Results	Avg Number of Textual Fields	Avg Number of Empty Textual Fields	Avg Number of Words per Result
Fashion	1120	33	1	49
Hotel Supplies	2210	17	8	96
Design	1713	32	3	69

Methodology - Retrieval System

1. Baseline system that combines BM25 with BGE-M3 embeddings
2. Created indexes for each dataset
3. Retrieved results based on a predefined list of queries
    1. Derived from real traffic data or
    2. Carefully crafted by human experts
4. Aim was to include queries and results that included hits and misses generated by both lexical and semantic retrievers

Methodology - Relevance Judgement Strategy

1. Two expert human annotators assessed the relevance of each pair on a 0-2 rating scale:
   - 2: Highly relevant, a perfect match for the query
   - 1: Somewhat relevant, a result that partially matches the query's intent
   - 0: Not relevant, a poor result that should not be shown
2. Grading guidelines were adapted to suit the specific characteristics of the datasets.

Query: *"v-neck white tee"*

Image	Search Result	Relevance Judgment
	prod_name: Premium ELKE vneck tee, index_name: Ladieswear, detail_desc: V-neck T-shirt in airy slub lin[...], department_name: Jersey/Knitwear Premium, index_group_name: Ladieswear, colour_group_name: White, product_type_name: T-shirt, graphical_appearance_name: Solid, perceived_colour_value_name: Light, perceived_colour_master_name: White	2

Methodology - Inter-Annotator Agreement

1. We use Cohen's Kappa to assess reliability of relevance judgements
    1. Used to quantify inter-annotator agreement for categorical data
    2. Ranges from -1 to 1, where 1 indicates strong agreement, while values closer to 0 suggest agreement no better than chance
2. We calculate:
    1. Agreement between human annotators and LLM-generated annotations
    2. Agreement between the pair of human annotators

Cohen's kappa	Interpretation
0 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Methodology - Models

1. Range of LLMs and MLLMs with varying levels of performance and cost
2. OpenAI Models:
   1. GPT-4V (gpt-4-vision-preview)
   2. GPT-4o (gpt-4o-2024-05-13)
   3. GPT-4o-mini (gpt- 4o-mini-2024-07-18)
3. Anthropic Models:
   1. Claude 3.5 Sonnet
   2. Claude 3 Haiku

Methodology - Prompts

1. Prompt guides the model to generate accurate relevance judgements
2. In multimodal setup, prompt references and includes the image
3. Prompt instructs the model to provide an explanation for its relevance judgement

Haiku's Prompt Template (Text-only Setup)

```
You are an assistant responsible for rating how the retrieved result is relevant to
the query. Output a token: "2", "1", or "0" followed by a full explanation.
Guidelines:
"2" - The result matches exactly with what the user's query is looking for.
"1" - The result is not exactly with what the user's query is looking for. But it's
pretty similar. As our aim is to be strict on exact matches, this grade is less
likely to be used.
"0" - The result is not related to the query at all.

Result: {{document}}
Query: {{query}}
Output:"
```

Haiku and Sonnet's Prompt Template (Multimodal Setup)

```
You are an assistant responsible for rating how the retrieved result is relevant to the query. If an image
is available, use it to determine the relevance to the query. Output a token: "2", "1", or "0" followed by
a full explanation.
Guidelines:
"2" - The result matches exactly with what the user's query is looking for.
"1" - The result is not exactly with what the user's query is looking for. But it's pretty similar. As our
aim is to be strict on exact matches, this grade is less likely to be used.
"0" - The result is not related to the query at all.

Result: {{document}}
Query: {{query}}
```

```
{{image}}
```

```
Token:
```

Results - Multimodal vs Single-modality Evaluation

1. LLM performance is dependent on the use case
2. No single model outperforms all the others across every use case
3. Tailoring a model's prompt to a specific domain can help
4. Vision component helps in larger models, but not the smaller models

	GPT-4v		GPT-4o		GPT-4o mini		Sonnet		Haiku		Human
	MM	Text	MM	Text	MM	Text	MM	Text	MM	Text	MM
Fashion	0.503	0.498	**0.613**	0.606	0.424	0.382	0.441	0.387	0.371	0.431	0.680
Hotel Supplies	0.620	0.596	0.627	0.582	0.506	0.565	0.634	**0.638**	0.471	0.560	0.641
Design	0.320	0.317	**0.404**	0.331	0.294	0.299	0.351	0.381	0.260	0.309	0.447
Average	0.481	0.471	**0.548**	0.506	0.408	0.415	0.475	0.469	0.368	0.433	0.589

Results - Cost-Accuracy Trade-off

1. GPT-4V is the most expensive model with high costs for tokens and image processing but performs strongly on both text and multimodal tasks
2. GPT-4o offers higher performances at a lower cost compared to GPT-4V
3. Sonnet is cheaper than GPT-4o but the performance also suffers.
4. Haiku is a very good choice for low-budget tasks

	GPT-4V	GPT-4o	GPT-4o-mini	Sonnet	Haiku
$/1M Input tokens	10.00	5.00	0.15	3.00	0.25
$/1M Output tokens	30.00	15.00	0.60	15.00	1.25
$/1M images (low resolution)	425.00	425.00	425.00	1048.58	87.38

Results - Prompt Engineering

1. Strict guidelines - results improved after we included instructions to prefer grades 2 (GREAT) and 0 (BAD)
2. Smaller models are more sensitive to prompt complexity
3. Prompts are model specific
4. Asking for explanations helps. It also helps us improve the prompt iteratively.

Conclusion

1. We have presented a new analysis of MLLMs-as-a-Judge, to assess the cost-accuracy trade-offs of relevance judgement capabilities of MLLMs
2. Various LLMs have shown potential, but no single LLM showed optimal cost-accuracy across all use cases evaluated
3. Choosing the best LLM judge for a given use-case is time-intensive and financially-demanding
4. We would like to encourage future work in the following directions:
    1. Improving the abilities of general MLLMs across use cases
    2. Improving cost and computational efficiency of large MLLMs
    3. Creating small MLLMs that are optimized for judging relevance in cost-optimal ways for more specialized applications