DispatchQA: A Benchmark for Small Function Calling Language Models in E-Commerce Applications

Joachim Daiber, Victor Maricato, Ayan Sinha, and Andrew Rabinovich

OVERVIEW

- **DispatchQA** measures how well small language models (SLMs) perform function calling for e-commerce search.
- Includes strong, replicable baselines based on fine-tuned Llama 3.1 8B [2].
- Code & models at github.com/upwork/dispatchqa

MOTIVATION

- Conversational, informal language can be challenging for traditional search systems.
- *Function calling* translates natural language queries into structured API calls.
- LLM APIs too slow for latency-sensitive production search systems.
- SLMs promising alternative, but need to evaluate for search quality and latency.

FINDINGS

- Query understanding as function calling improves search quality.
- **SLM Quality & Speed**: Fine-tuned Llama 3.1 8B matches GPT-4o quality at lower latency.
- **Structured Decoding**: Structured decoding implementation and instructions (e.g. repeat schema in prompt?) are critical for fast and accurate search via function calling.

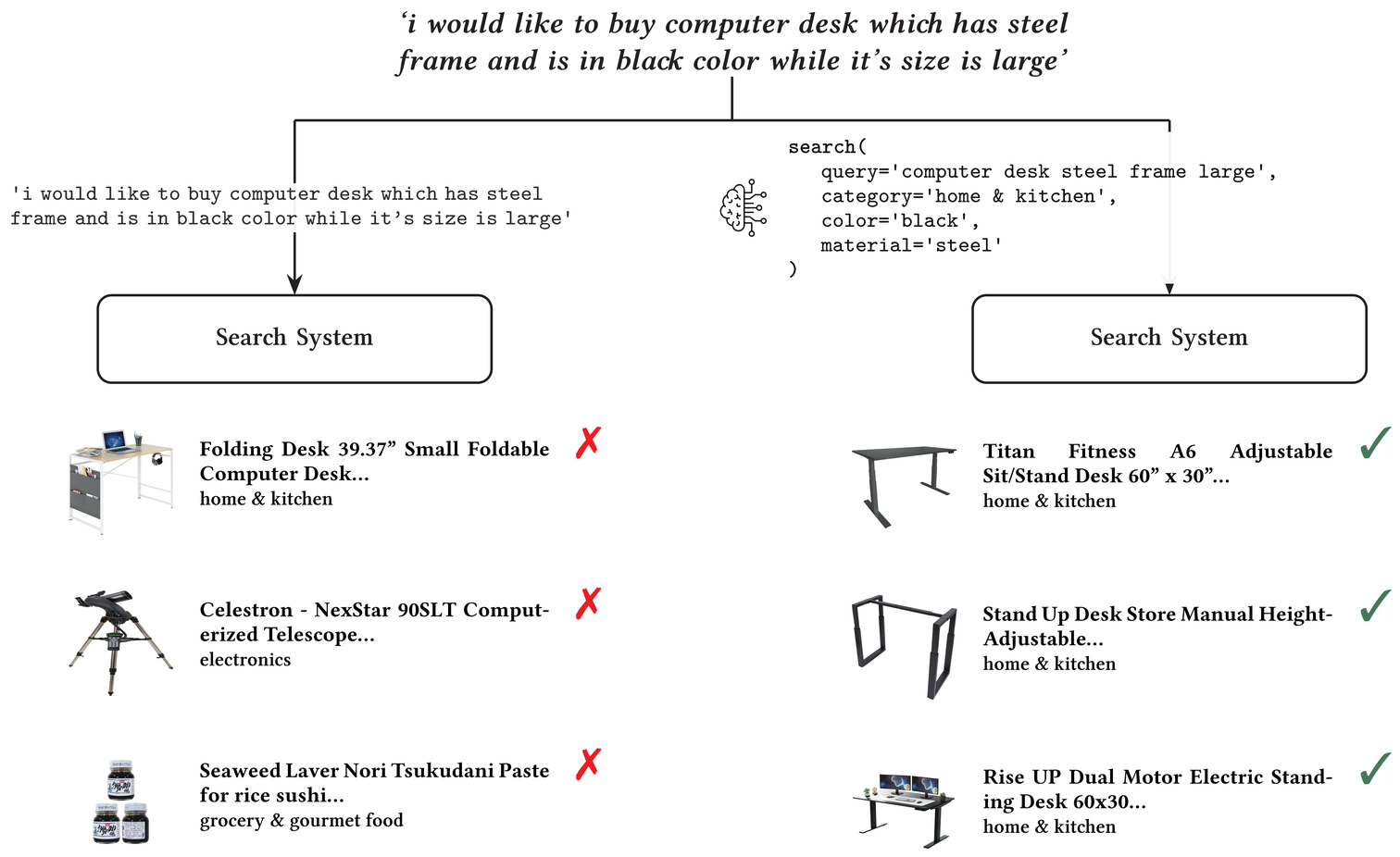

Figure: A baseline search using only the raw query is compared against a search using a query understanding model that generates a structured API call.

DISPATCHQA BENCHMARK

- Application-based evaluation: Search with and without SLM-based query understanding.
- Search quality via LLM-as-a-judge
  - GPT-4o judge evaluates relevance of search results
  - Validated against human judges on WANDS [1] and WebShop [3], see paper for details
- Measure both search quality and end-to-end latency.
- DispatchQA 1.0: Queries and products from WebShop [3].

SEARCH EXPERIMENTS

	Search quality metrics					Latency
System	NDCG@10	MAP@10	MRR@10	R@10	P@10	P50 in ms
Baseline	0.24	0.18	0.19	0.38	0.13	**8**
Closed models (GPT-4 family)						
GPT-4o	0.32	0.24	0.27	0.49	0.18	593
GPT-4o mini	0.32	0.25	0.26	0.51	0.19	672
GPT-4.1 nano	0.37	**0.30**	**0.31**	0.54	0.22	496
Closed reasoning models (GPT-5)						
GPT-5	**0.37**	0.29	0.31	0.55	**0.24**	11470
GPT-5 mini	0.35	0.28	0.31	0.50	0.22	6836
GPT-5 nano	0.36	0.29	0.31	0.55	0.21	5576
Small local models						
Llama 3.1 8B Instruct (SGLang)	0.32	0.24	0.27	0.50	0.20	284
Llama 3.1 8B Instruct Fine-tune (SGLang)	0.33	0.25	0.26	0.51	0.21	313
Llama 3.1 8B Instruct LoRA (SGLang)	0.35	0.26	0.27	**0.57**	0.20	321
DeepSeek R1 Distill Qwen 1.5B (SGLang)	0.24	0.18	0.20	0.40	0.13	694
Qwen3 0.6B (SGLang)	0.25	0.19	0.20	0.41	0.14	586
Gemma 3 4B IT (SGLang)	0.33	0.26	0.28	0.52	0.21	495

Table: DispatchQA comparison of closed LLMs and smaller, specialized models.

REFERENCES

[1] Y. Chen, S. Liu, Z. Liu, W. Sun, L. Baltrunas, and B. Schroeder. Wands: Dataset for product search relevance assessment. In Proceedings of the 44th European Conference on Information Retrieval, 2022.

[2] A. D. et al. The Llama 3 herd of models. CoRR, abs/2407.21783, 2024.

[3] S. Yao, H. Chen, J. Yang, and K. Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 20744–20757. Curran Associates, Inc., 2022.