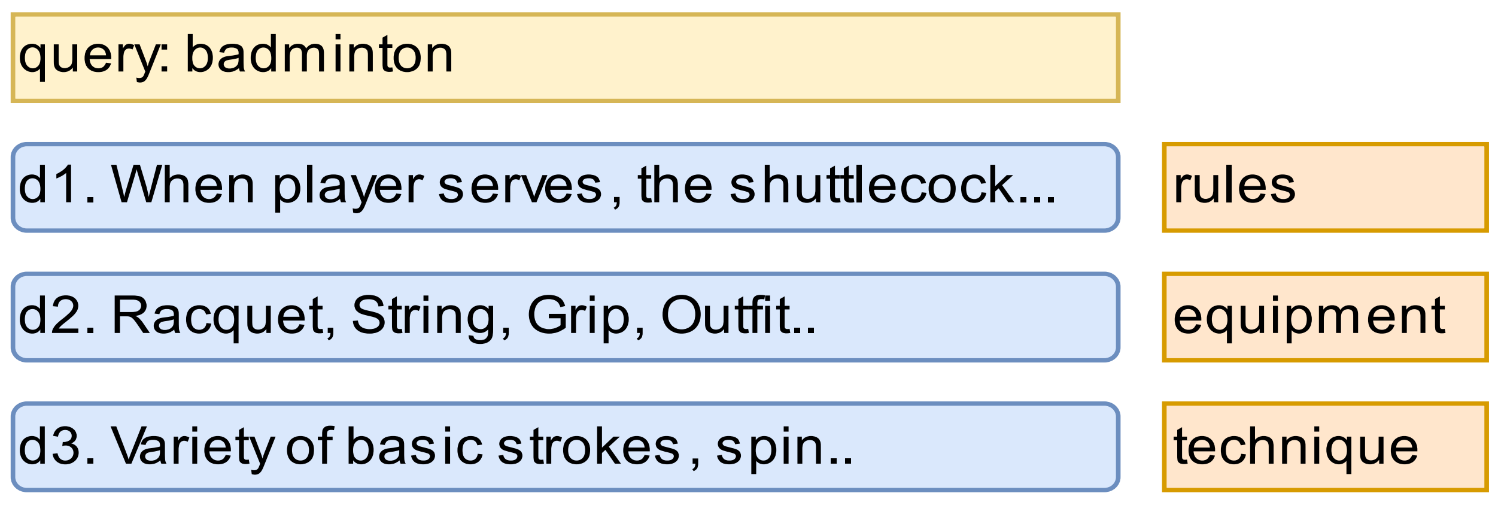
Generating Search Explanations using Large Language Models

Arif Laksito and Mark Stevenson
*S*chool of Computer Science, University of Sheffield, Sheffield, United Kingdom

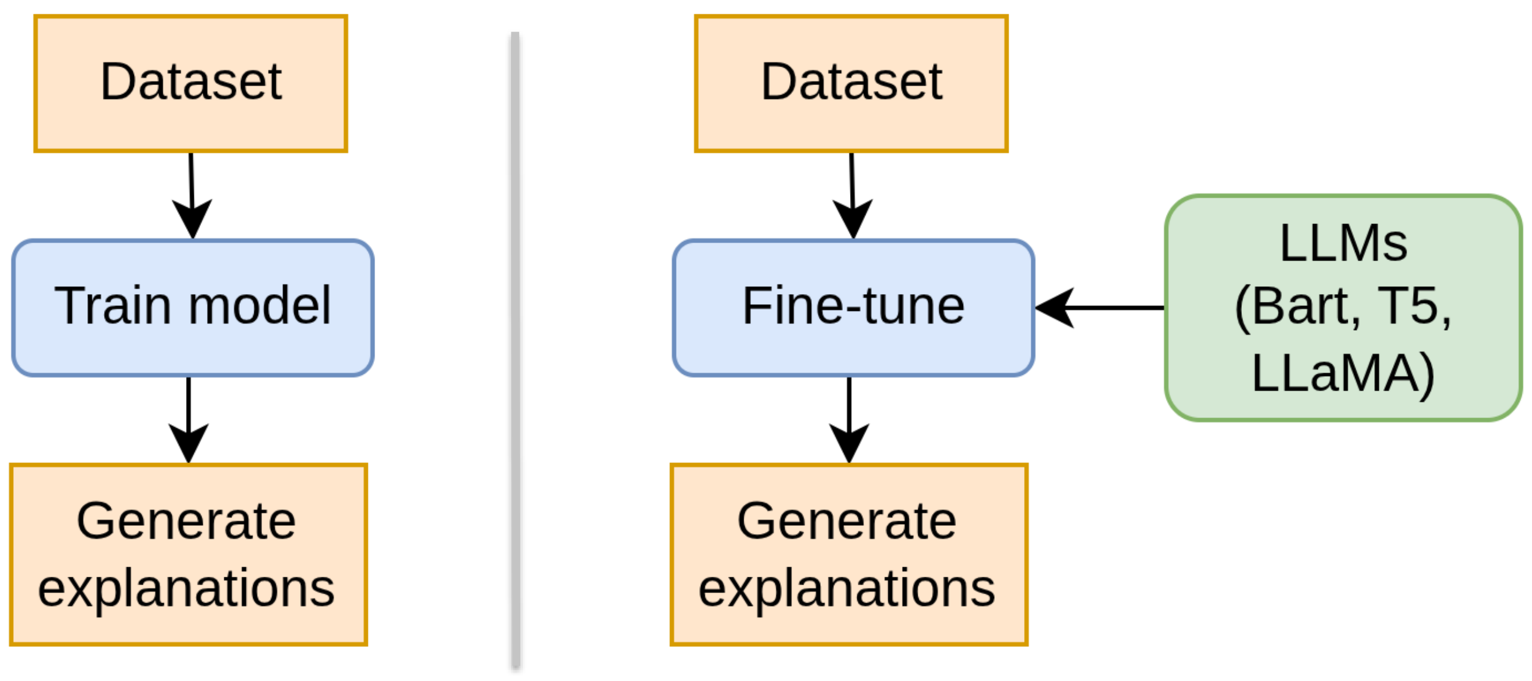
Motivation

- In search systems, users tend to submit under-specified queries with multiple potential intents.
- Traditional document snippets help users scan result quickly but often fail to explain why documents are relevant.
- There is a growing need for concise explanations to clarify document relevance.

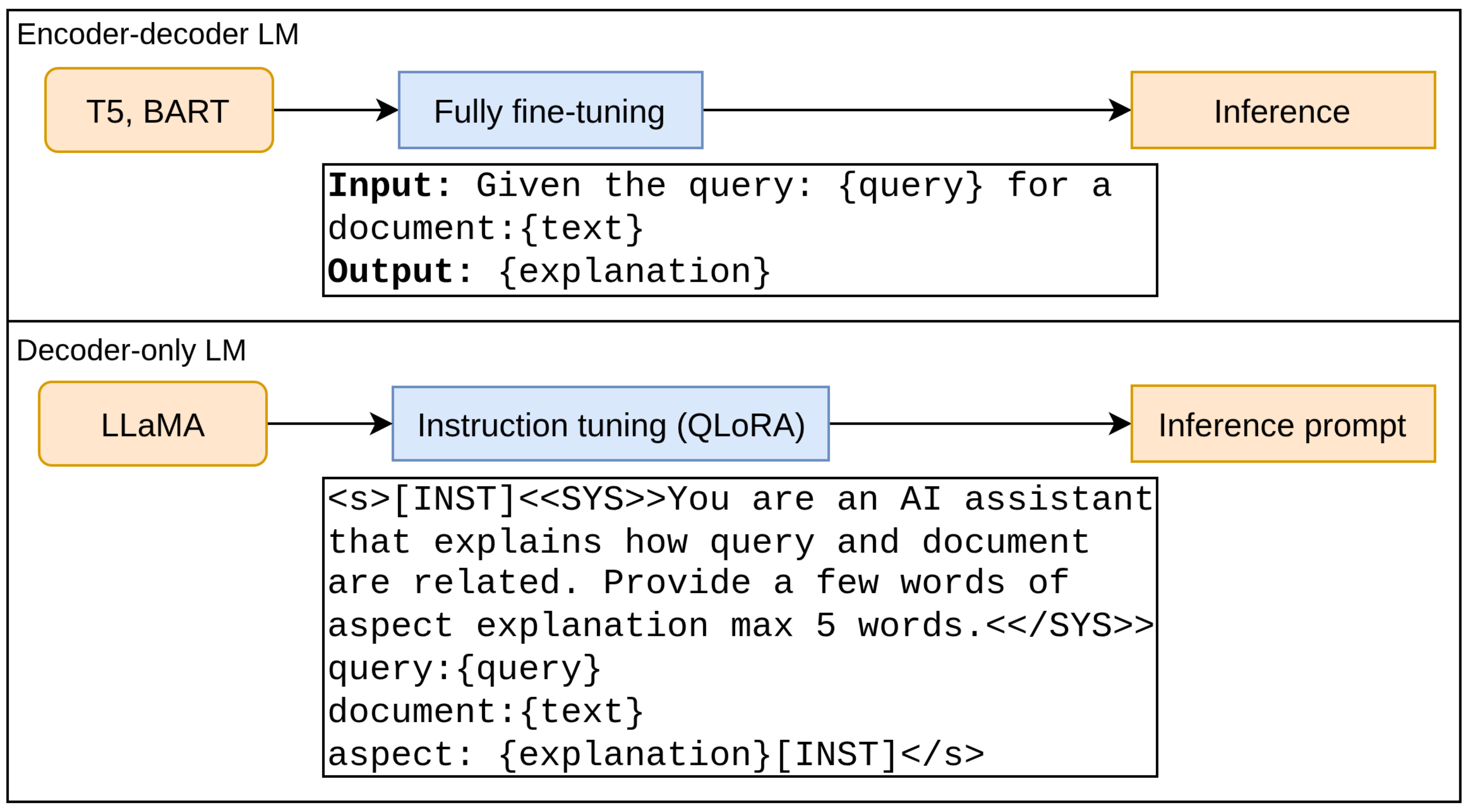

query: badminton
d1. When player serves, the shuttlecock...
rules
d2. Racquet, String, Grip, Outfit..
equipment
d3. Variety of basic strokes, spin..
technique

Example of concise explanations in search system

- Our contribution is leveraging LLMs to generate explanations in search results.



Previous research (left) trained transformer models from scratch to generate explanations. In contrast, we fine-tune pre-trained LLMs (right).



```
Input: Given the query: {query} for a
document:{text}
Output: {explanation}
```

```
<s>[INST]<<SYS>>You are an AI assistant
that explains how query and document
are related. Provide a few words of
aspect explanation max 5 words.<</SYS>>
query:{query}
document:{text}
aspect: {explanation}[INST]</s>
```

Fine-tuning strategies for encoder-decoder and decoder-only language models.

Dataset

- Constructed Wikipedia articles titles as queries and their section headings as explanations.
- Consists of 54k samples with 19k queries.
- Split into 40k train, 4k dev and 10k test.

Fine-tuning LLMs

- Fine-tune both encoder-decoder (T5, BART) and decoder-only models (LLaMA).
- Employ a natural language input representation for training encoder-decoder models.
- Adopt an instruction-tuning framework where inputs are framed as natural prompts followed by expected outputs for decoder-only models.

Results

	Architecture	Parameters	METEOR	ROUGE-1	BERTScore
Transformer	Encoder-decoder	21M	0.0747	0.1264	0.3057
Bert2Bert	Encoder-decoder	247M	0.0846	0.1323	0.2970
Bert2Gpt	Encoder-decoder	262M	0.1158	0.1917	0.3157
0-shot Llama(v2)	Decoder-only	13B	0.0920	0.1145	0.1830
0-shot Llama(v3)	Decoder-only	70B	0.1215	0.1813	0.2920
FT BART	Encoder-decoder	139M	0.2331	0.3923	0.4771
FT T5	Encoder-decoder	220M	0.2723	0.4301	0.5202
FT Llama(v2)	Decoder-only	13B	0.2759	0.3896	0.4362
FT Llama(v3)	Decoder-only	70B	**0.3222**	**0.4993**	**0.5652**

- All fine-tuned models significantly outperform zero-shot and baseline encoder-decoder models across overall metrics.
- Fine-tuned LLaMA models show substantial gains over zero-shot settings, indicating that LLaMA benefits from fine-tuning for this task.

Examples of generated text

query	Érick Valencia Salazar
document	Once the clashes were over Valencia was transferred to the PGRs installations in Mexico City..
reference text	Imprisonment
Transformer	Background
0-shot LLaMA (v3)	Legal status
FT LLaMA (v3)	Arrest and imprisonment
FT BART	Arrest
FT T5	Arrest

Comparison of generated explanations from several models.

Conclusion

Fine-tuned LLMs can produce concise and plausible aspect-based explanations in search system.