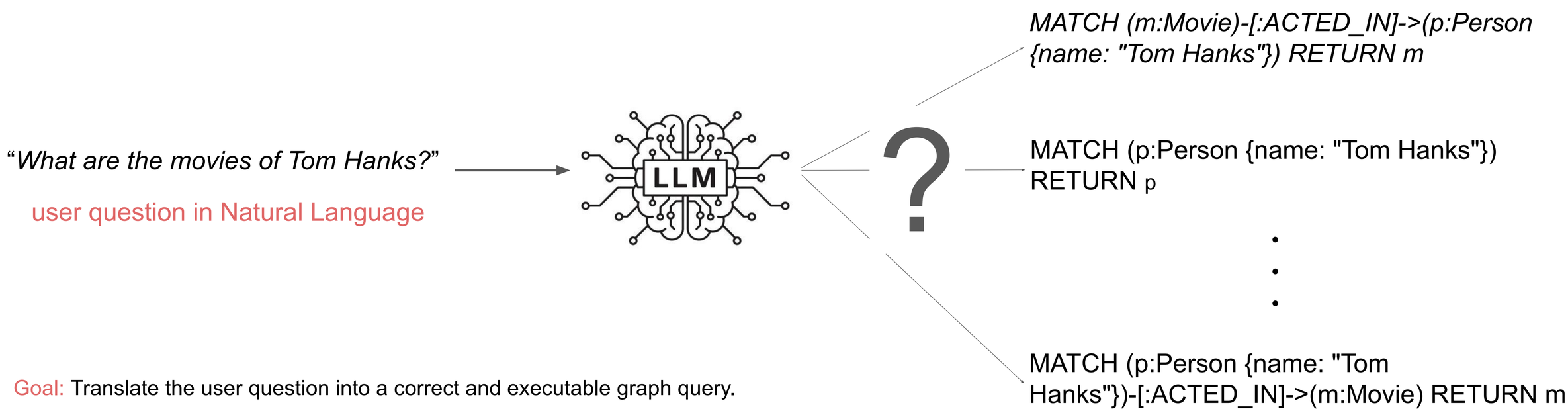
Improving Text2Cypher with Confidence-Based Test-Time Strategies

Rima Dessi'[1], Makbule Gulcin Ozsoy[2]
[1]Higher Colleges of Technology, Sharjah, UAE, [2]Neo4j, London, UK
rdessi@hct.ac.ae, makbule.ozsoy@neo4j.com

Motivation

Text2Cypher



Goal: Translate the user question into a correct and executable graph query.

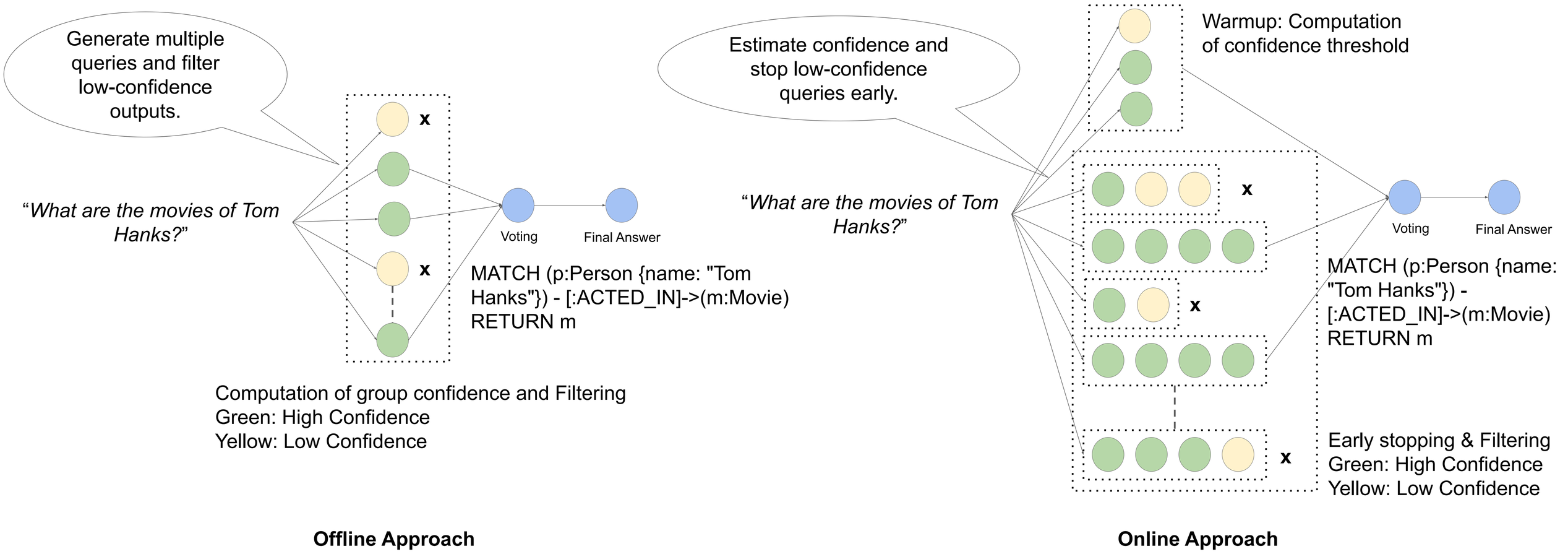
Challenge: LLMs generate multiple candidate queries (traces), but not all are correct or executable.

Proposed idea: Using confidence-based filtering to select the most reliable query.

Picture: https://stock.adobe.com/ae/search?k=llm

Confidence-aware strategies with offline and online approaches

Architecture



Results

Model type	Inference Mode	ROUGE-L	Execution Success Ratio
Reasoning model	Base Online Offline	0.4025 0.4919 0.4888	0.1267 0.1622 0.1445
Instruction-tuned	Base Online Offline	0.7004 0.7060 0.7095	0.8200 0.8276 0.8416

Table1:Comparison of reasoning (DeepSeek-R1-Distill-Qwen-7B) and instruction-tuned (Gemma2-9B-it) models

Model	Diversity	Inference	ROUGE-L (lexical)	ROUGE-L (exec)	Execution Success
Gemma-2-9b-it	Light	Base Online Offline	0.7004 0.7060 (+0.56) 0.7095 (+0.91)	0.2343 0.2399 (+0.56) 0.2427 (+0.84)	0.8200 0.8276 (+0.76) 0.8416 (+2.16)
Gemma-2-9b-it	Moderate	Base Online Offline	0.6817 0.7162 (+3.45) 0.7081 (+2.64)	0.1982 0.2411 (+4.29) 0.2375 (+3.93)	0.7769 0.8530 (+7.61) 0.8162 (+3.93)
Gemma-2-9b-it	High	Base Online Offline	0.6286 0.7099 (+8.13) 0.6948 (+6.62)	0.1680 0.2396 (+7.16) 0.2224 (+5.44)	0.6388 0.8365 **(+19.77)** 0.7503 (+11.15)
Qwen2.5-7B-Instruct	Moderate	Base Online Offline	0.6898 0.7125 (+2.27) 0.7202 (+3.04)	0.1917 0.2189 (+2.72) 0.2391 (+4.74)	0.7098 0.7833 (+7.35) 0.7896 (+7.98)

Table2: Performance across diversity levels (% improvement over base)