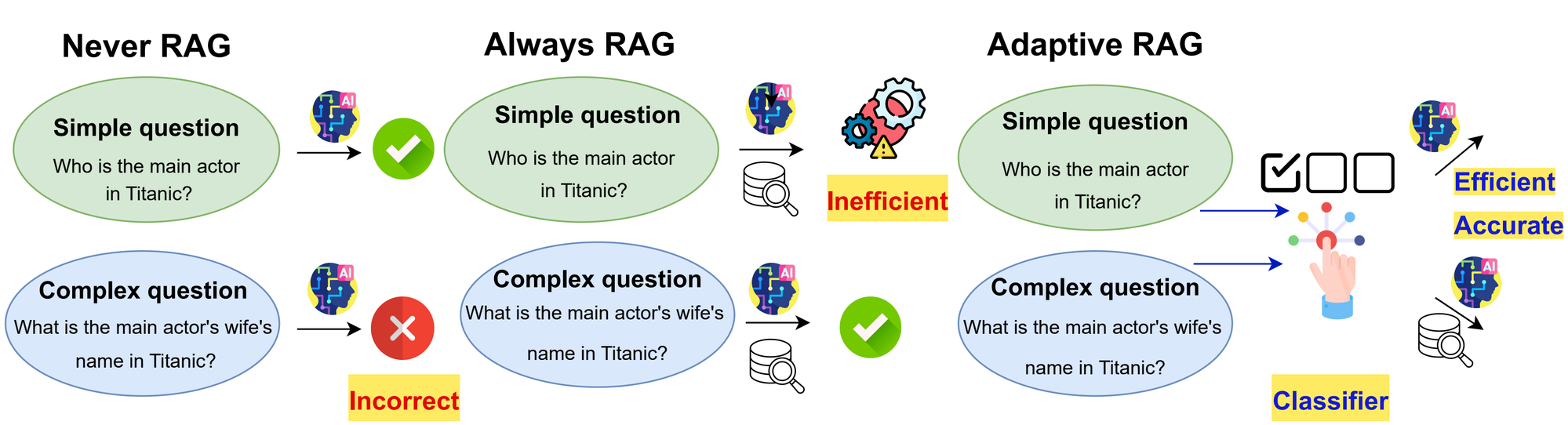
LLM-Independent Adaptive RAG: Let the Question Speak for Itself

Maria Marina[1,2] Nikolay Ivanov[2] Sergey Pletenev[1,2] Mikhail Salnikov[1,2] Daria Galimzianova[3,4]
Nikita Krayko[3] Vasily Konovalov[1,2,5] Alexander Panchenko[2,1] Viktor Moskvoretskii[6]

Skoltech [1]AIRI [2]Skoltech [3]MWS AI [4]MBZUAI [5]MIPT [6]EPFL

EMNLP 2025



Setting

- Many RAG systems are heavy (based on internal representations), or they introduce misleading context even when the LLM could answer without it
- We introduce a **lightweight adaptive RAG pipeline based on LLM-independent features** that triggers retrieval only when needed, avoiding the risk of noisy context
- Our solution builds on the LLaMA-3.1-8B-Instruct model and uses **7 groups of** lightweight external **features (27 in total)**, evaluated across **6 datasets**.

Lightweight LLM-independent features

Popularity of the entity in the question, estimated through three proxy: *graph* (number of Wikidata triples containing the entity), *popularity* (page views of the Wikipedia page of the entity) and *frequency* (occurrences in the reference text collection).

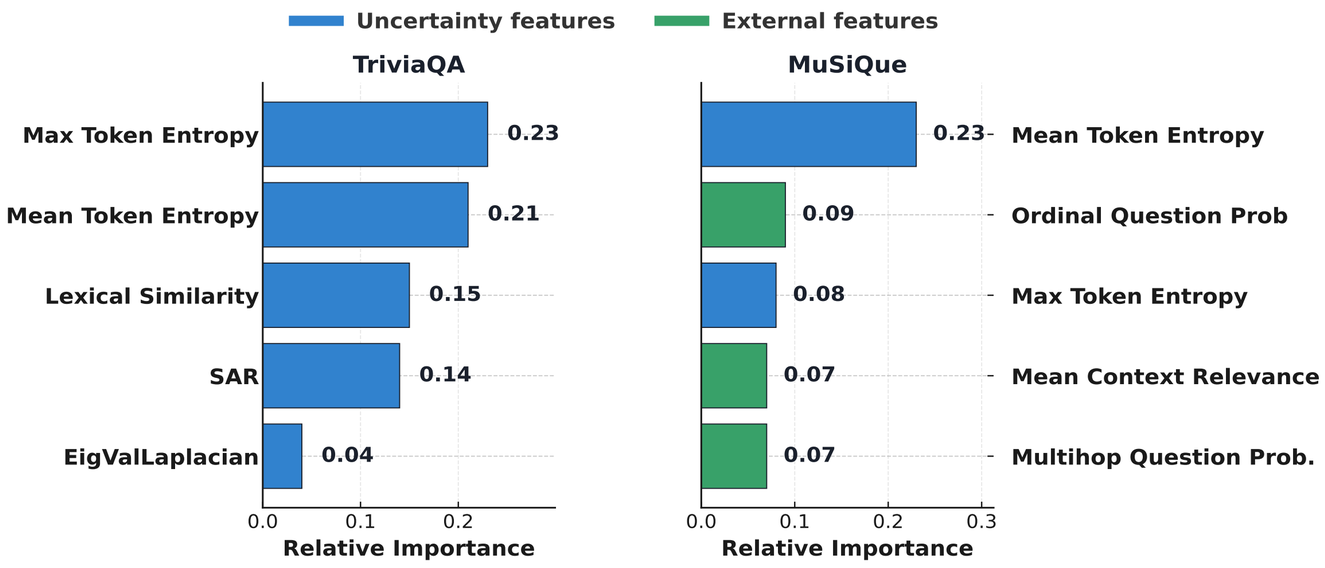
Question type - one of the 9: ordinal, count, generic, superlative, difference, intersection, multihop, comparative, and yes/no

Question complexity reflects the difficulty of a question, considering the reasoning steps required

Context relevance - each question–context pair is fed to a BERT-base-uncased cross-encoder model

Knowledgeability features assign a score to each entity, reflecting the LLM's verbalized uncertainty about its knowledge (possible to precompute!)

Feature importances



Results

- Proposed LLM-independent adaptive RAG approach shows results comparable even to multistep approaches
- *Can external features replace uncertainty?* At at least one external feature performs comparably to UE-based methods
- *Can external features complement uncertainty?* On some datasets (MuSiQue) we see that external features have some additional signal

Method	TriviaQA			MuSiQue			Avg
	InAcc↑	LMC↓	RC↓	InAcc↑	LMC↓	RC↓	InAcc
Never RAG	63.6	1.0	0.00	10.6	1.0	0.00	32.8
Always RAG	61.0	1.0	1.00	10.0	1.0	1.00	38.4
Multi-Step Adaptive Retrieval							
AdaptiveRAG	62.8	1.5	0.54	**14.0**	3.6	3.63	40.3
DRAGIN	**66.6**	4.1	2.06	13.4	6.3	3.15	41.1
FLARE	64.8	2.1	1.39	9.0	4.1	3.10	37.0
Rowen$_{CM}$	65.6	28.7	7.12	10.4	42.1	9.52	37.5
Seakr	65.6	14.6	1.00	11.8	12.3	2.40	37.8
Uncertainty Estimation							
EigValLaplacian	64.4	1.3	0.34	10.0	2.0	0.96	38.7
MaxTokenEntropy	63.4	1.3	0.31	11.2	1.7	0.72	38.4
Hybrid UE	63.8	1.3	0.27	11.0	1.7	0.74	39.3
External Features							
Graph	63.6	1.0	0.32	10.0	1.0	1.00	38.4
Popularity	63.0	1.0	0.16	10.6	1.0	0.89	38.5
Frequency	63.2	1.0	0.04	10.4	1.0	0.90	38.7
Knowledgability	63.0	1.0	0.28	10.2	1.0	0.46	38.9
Question type	**64.0**	1.0	0.29	10.4	1.0	0.90	38.4
Question complexity	63.6	1.0	0.00	10.6	1.0	0.95	38.8
Context relevance	62.6	1.0	1.00	11.0	1.0	1.00	38.2
Hybrids with External Features							
Hybrid$_{\neg UFP}$	63.4	1.1	1.0	10.6	1.2	1.0	38.3
Hybrid$_{External}$	63.2	0.2	1.0	10.6	2.0	1.0	37.9
Hybrids with Uncertainty and External Features							
Hybrid$_{\neg FP}$	64.6	1.3	1.0	**12.2**	1.4	1.0	39.3
All	63.2	1.3	1.0	11.2	1.1	1.0	38.1
Ideal	**73.6**	1.4	0.36	**16.4**	1.9	0.89	47.1

m.marina.scientia@gmail.com A.Panchenko@skol.tech