NAVER LABS

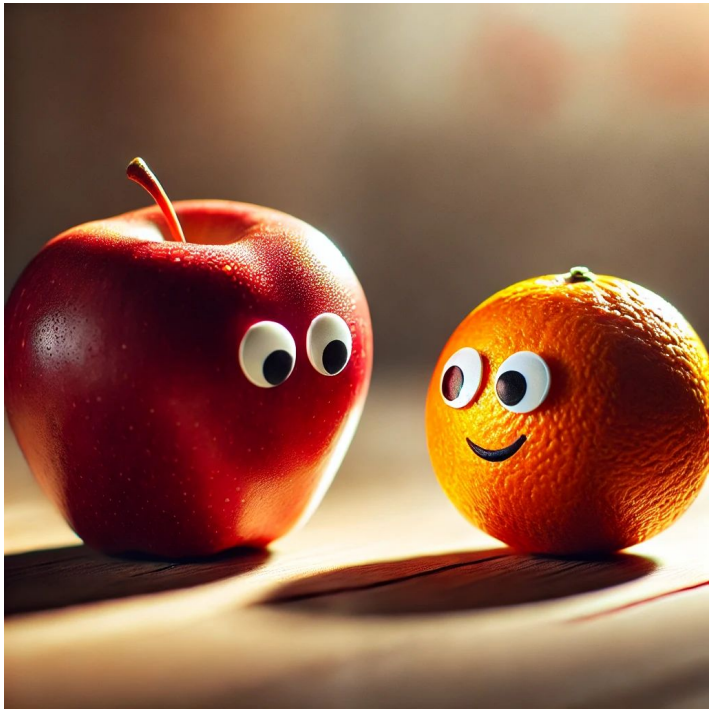
BERGEN

BEnchmarking Retrieval-augmented GENeration

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Vassilina Nikoulina, Stéphane Clinchant

Motivation: compare apples and oranges

- **RAG experimental setups are fragmented:**
  - Reviewed 20+ papers
  - different datasets, variety of retrievers, metrics, top-k documents
  - Different settings: zero-shot, few-shot, etc.
- **What's a good baseline for RAG?**



BERGEN:

An open source library to ease reproducibility of RAG experiments

A set of **recommendations** for strong baselines in RAG:

- Evaluation metrics
- Datasets
- Retrieval Setup

An open source library to ease reproducibility

- Support for various retrievers (20+), rerankers (4+), and LLMs (20+)
- Comprehensive evaluation metrics (Match, EM, LLMs, ...)
- Support for multilingual experiments

Welcoming contributors! 🤗

https://github.com/naver/bergen

```
python bergen.py retriever='splade-v3'
reranker='minilm6'
generator='SOLAR-10.7B'
dataset='kilt_nq' train='lora'
```

- **Evaluation Datasets**: NQ, TriviaQA, HotPot QA, Wizard of Wikipedia, ELI5, WikiQA, TruthfulQA, PopQA, ASQA, SCIQ, MKQA, XorTidyQA
- **Collections**: KILT Wikipedia passages ~21*M,* Multilingual Wikipedia, Pubmed …
- **Setting**: Focus on **Zero-Shot RAG**
- **Context**: Top-5 documents

Comparison of metrics

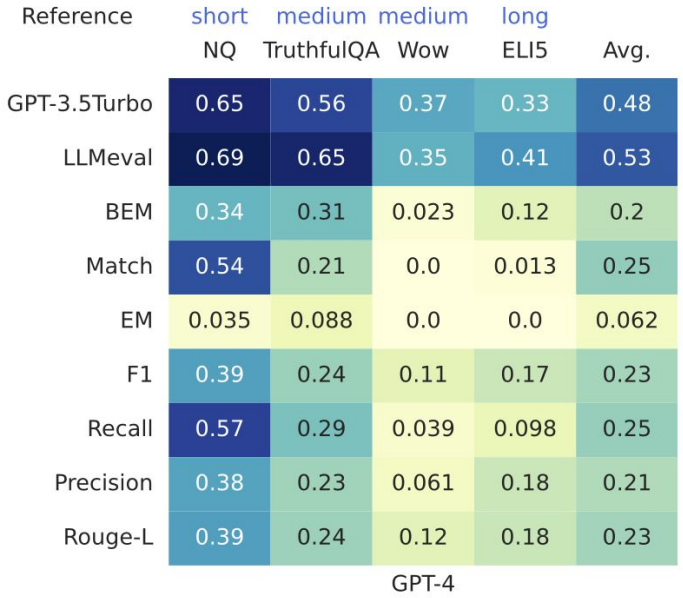


Figure 2: Correlation of different metrics with GPT-4-as-a-judge for datasets with varying reference label lengths (short, medium, and long).

RQ1: *Which metrics are most effective for evaluating open-ended text generation and comparing RAG systems?*

LLMeval: LLM-as-a-Judge based on SOLAR-10.7B

*LLMeval **closely aligns with GPT-4's evaluation**, followed by **Match and Recall**, making them the most effective non-commercial metrics for (zero-shot) RAG evaluation, among the ones tested.*

Figure 3: Performance gain w/ and w/o retrieval (SPLADE-v3 + reranking (RR) with DeBERTa-v3) on different datasets with SOLAR-10.7B.

RQ2: *Which **datasets** are **most suitable** for RAG*

We find:

- ASQA
- Hotpot QA
- NQ
- TriviaQA
- POPQA

Not useful: KILT ELI5!, WoW

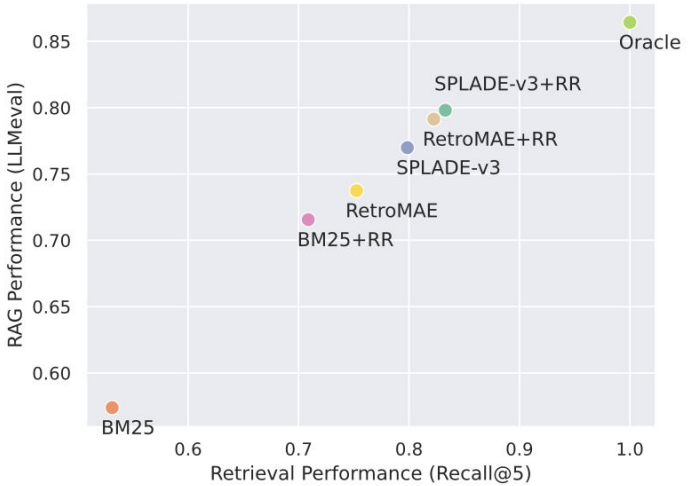


Figure 4: Impact of retrieval performance on RAG Performance for SOLAR-10.7B on NQ with different ranking systems. RR means with additional re-ranking using DeBERTa-v3.

RQ3: *Does **retrieval quality positively** impact **generation quality**?*

Good performing RAG pipeline should rely on **strong retrievers**

Reranking is often overlooked in RAG papers and should be used for strong baselines

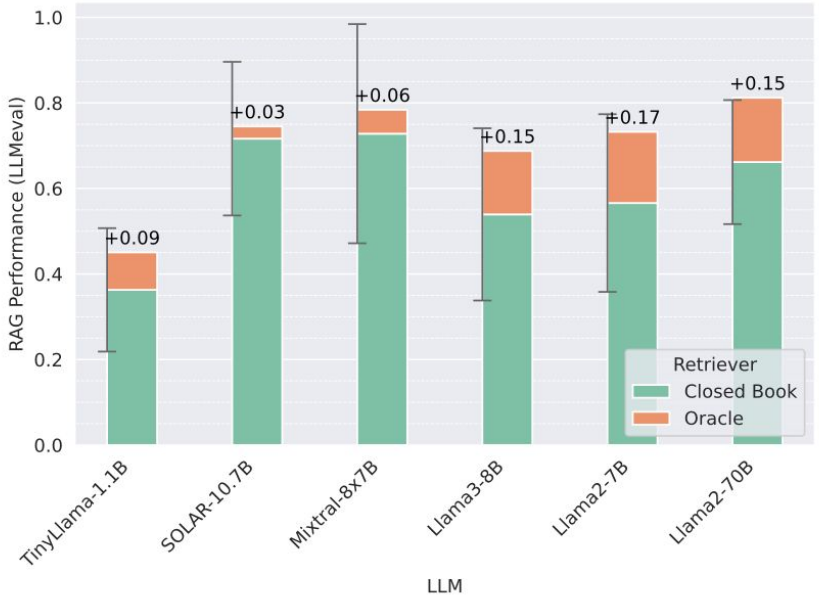


Figure 5: Performance gains w/ and w/o oracle retrieval for LLMs with different sizes. Comparing closed book *vs* oracle passages averaged over all QA datasets in KILT.

RQ4: *What is the impact of the **LLM size** in RAG?*

LLMs of all sizes benefit from retrieval augmentation

LLM	M	LLMeval
TinyLlama-1.1B-chat	0.56 (+0.13)	0.77 (+0.41)
Llama-2-7B-chat	0.64 (+0.03)	0.82 (+0.24)
Llama-3-8B-chat	0.66 (+0.02)	0.78 (+0.04)
SOLAR-10.7B	0.67 (-0.03)	0.84 (+0.05)
Mixtral-8x7B-inst.	0.68 (+0.01)	0.84 (+0.05)
Llama-2-70B-chat	0.69 (+0.04)	0.85 (+0.06)

Table 3: LLMs fine-tuned on NQ, for retrieval with SPLADE-v3 and reranking with DeBERTa-v3. Performance gains in absolute points compared to zero-shot is indicated in brackets.

RQ5: *How much performance can be gained by **fine-tuning**?*

All LLMs **benefit from FT**

Fine-tuning reduces performance gap between the smallest and the largest LLM, compared to zero-shot.

What is a good baseline?

Match Metric

Model	ASQA	NQ	TriviaQA	POPQA	HotPotQA
Llama-2-7B	68.4	61.6	87.9	60.2	45.9
Llama-2-70B	73.2	65.8	92.3	65.5	53.6
Mistral-8x7B	73.5	67.1	91.8	67.9	54.5
Solar-10.7B	76.2	70.2	92.8	71.2	53.9

top_k documents: 5

Retrieval: SPLADE-v3

Reranker: DeBERTa-v3

LLM: SOLAR-10.7B

Multilingual datasets:

- **XOR TydiQA**
- **MCQA**

Multilingual retriever/reranker:

- **BGE-m3**

Multilingual generator:

- **Command-R**

	en	ar	fi	ja	ko	ru
			MKQA			
No Ret	0.67	0.29	0.32	0.37	0.32	0.48
En Wiki	**0.76**	0.54	0.58	0.63	0.59	0.71
Multi Wiki	0.74	**0.57**	**0.64**	**0.64**	**0.62**	**0.72**
			XORQA			
No Ret	0.63	0.56	0.41	0.40	0.54	0.47
En Wiki	**0.73**	0.57	0.59	0.51	0.58	0.65
Multi Wiki	0.69	**0.70**	**0.74**	**0.62**	**0.66**	**0.74**

Multilingual collection provides richer multi-cultural knowledge and improves performance

Conclusion

BERGEN contains various metrics, datasets, models, search runs to build on.

More:

- Context Pruners (LLMLingua, RECOMP, etc)
- Query Generations
- Fine-tuning
- Multilingual RAG extensions

We are looking for contributors and suggestions to integrate new tasks, datasets etc.

https://github.com/naver/bergen

Recommendations:

- Datasets: ASQA, NQ, TriviaQA, PopQA
- Metrics: LLMEval + (Match | Recall)
- Models: Splade-v3 + deberta + SOLAR

Multilingual RAG:

- BGE-m3 + Command-R + multilingual Wiki + language-specific prompts