Answering Complex Open-ended Questions in the Era of Generative LLMs

Weronika Łajewska, Krisztian Balog
University of Stavanger, Norway

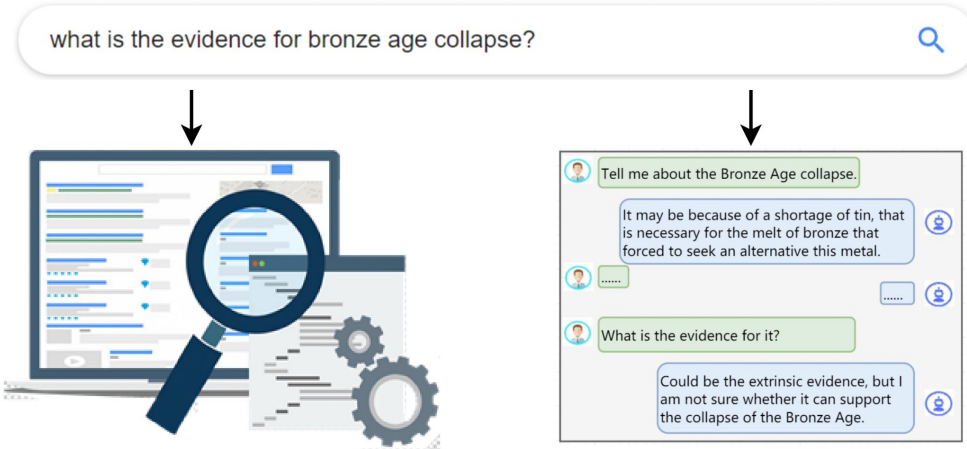
NorwAI Innovate, Sep 24, 2024

Contact information: weronika.lajewska@uis.no

Information Retrieval vs. Information Generation

Search engines

LLMs



Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Trans. Inf. Syst. 39, 4, Article 47 (October 2021), 29 pages. https://doi.org/10.1145/3432726

Information Retrieval vs. Information Generation

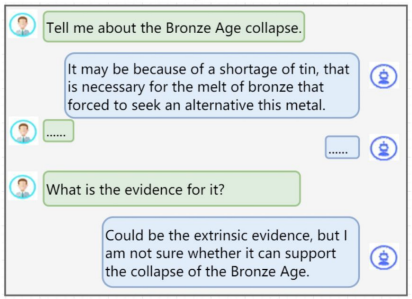
Search engines

→ Retrieve and rank existing web pages or documents based on relevance to the user's query

→ The sources of information is displayed directly to the user

→ Users often need to actively go through results

LLMs



Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Trans. Inf. Syst. 39, 4, Article 47 (October 2021), 29 pages. https://doi.org/10.1145/3432726

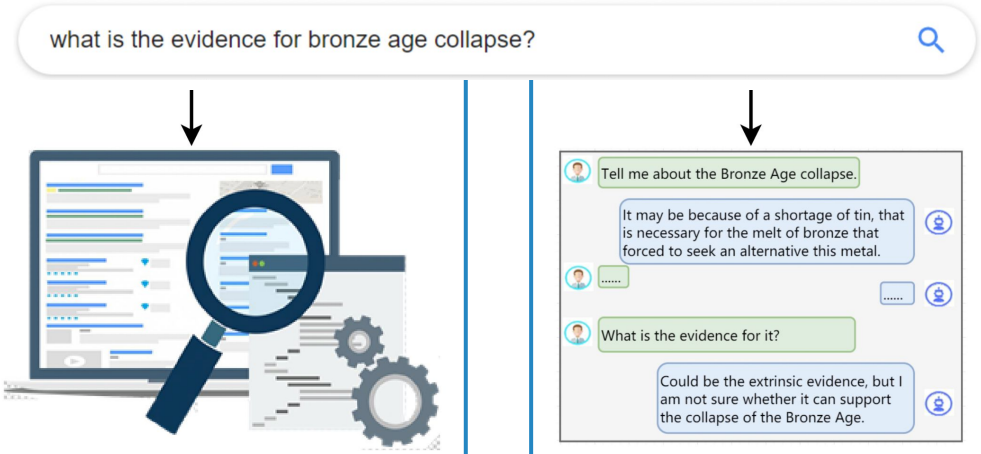
Information Retrieval vs. Information Generation

Search engines

→ Retrieve and rank existing web pages or documents based on relevance to the user's query

→ The sources of information is displayed directly to the user

→ Users often need to actively go through results



LLMs

→ Generate responses by processing the input query and synthesizing information from the vast amount of data they've been trained on

→ Responses do not explicitly cite the sources

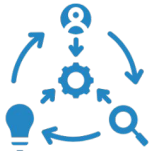
→ Immediate, cohesive answer is provided

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. ACM Trans. Inf. Syst. 39, 4, Article 47 (October 2021), 29 pages. https://doi.org/10.1145/3432726

Using LLMs to answer complex queries

Advantages

Response fluency and naturalness



Information synthesis

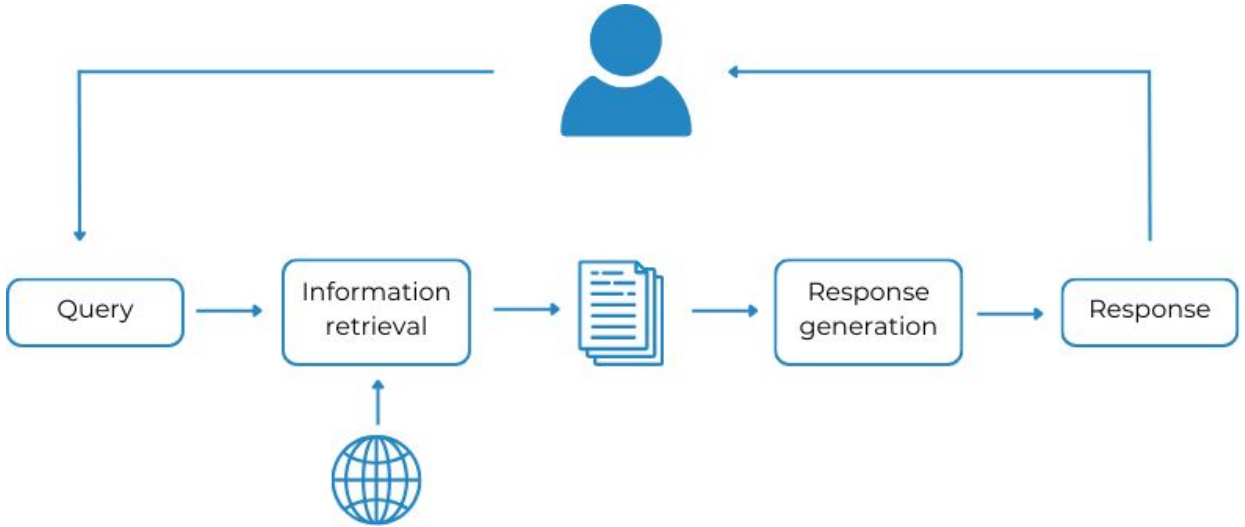
Issues

Factual correctness

Lack of transparency

No source attribution

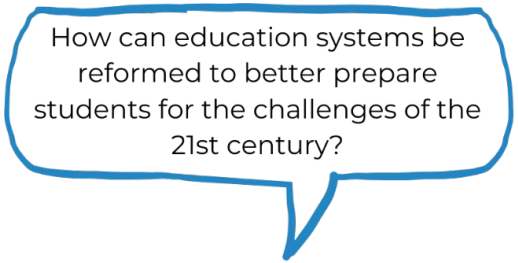
Retrieval-augmented generation (RAG)

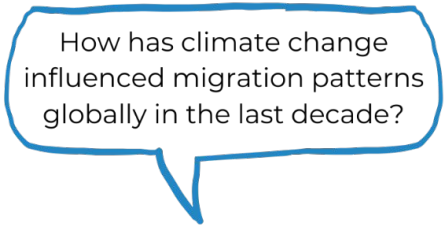


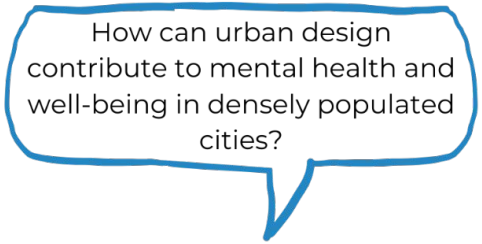
Challenges:

- Redundant information and overly long contexts can lead to the *“lost in the middle“* problem
- Response grounding is not guaranteed as some facts from the generated response may be not supported by the provided evidence

Complex open-ended questions







- They require covering multiple aspects or points of view
- The coverage of information in the response depends on user preferences, their background knowledge and previous interactions with the system
- In conversational setting, responses are expected to be short and concise
- There is a trade-off between response completeness and succinctness

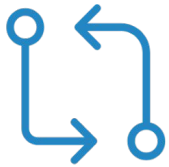
GINGER

Grounded **I**nformation **N**ugget-based
GEneration of Conversational
Information-Seeking **R**esponses

GINGER features

Ensuring grounding of the final response in the source passages to enable easy verification of source attribution

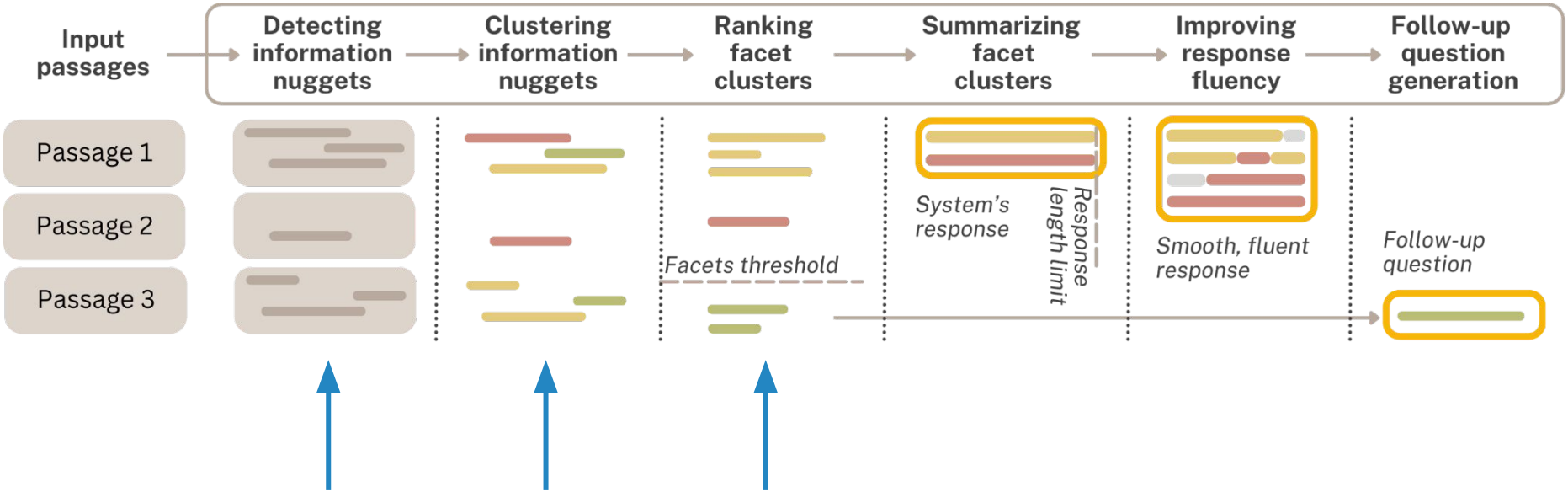
Controlling response completeness, by ensuring the coverage of a required number of query facets in the response within a predefined length limit



Suggesting relevant and answerable follow-up questions based on facets that could not be covered in the response due to length constraints

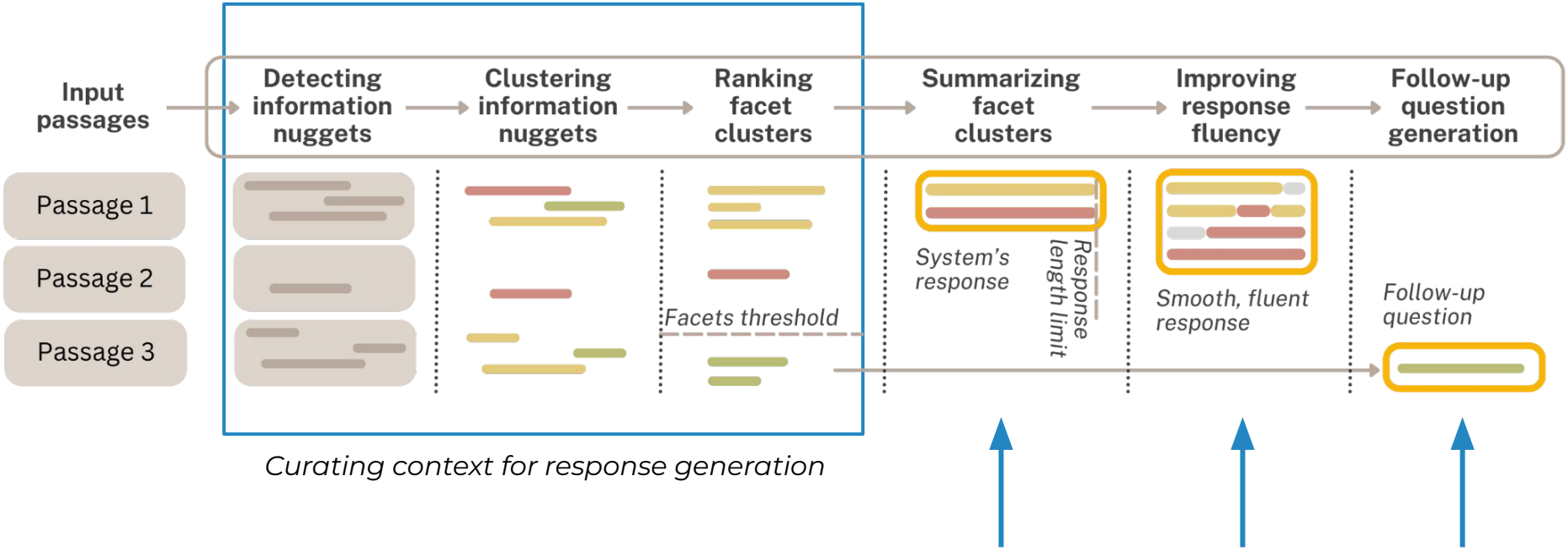
GINGER pipeline

Information nugget is a *"minimal, atomic units of relevant information"* in a passage



GINGER pipeline

Information nugget is a *"minimal, atomic units of relevant information"* in a passage



Does GINGER improve grounding and source attribution over the baseline?

- By operating on information nuggets, **GINGER grounds responses in specific facts and significantly improves source attribution over the baseline**

Method	Automatic evaluation			Human evaluation							
	Entailment	Contradiction	Completeness	Coh	Con	Eng	Fac	Suf	Pref	FQ_rel	FQ_use
Baseline	0.34	0.10	0.25	0.45	0.40	0.50	0.48	**0.52**	0.49	**2.58**	**2.63**
GINGER	**0.61***	**0.06**	**0.29**	**0.55**	**0.60***	**0.50**	**0.52**	0.48	**0.51**	**2.58**	2.60

Automatic and human evaluation of responses. Automatic measures target grounding (nugget entailment and contradiction) and completeness. Human evaluation reports the fraction of votes received when compared with the other method for (Coh)erence, (Con)ciseness, (Eng)agingness, (Fac)tuality, (Suf)ficiency, response (Pref)erence, and average scores for follow-up questions (on 3-point Likert scale) in terms of relevance (FQ_rel) and usefulness (FQ_use). Statistically significant differences ($p < 0.05$) with respect to the baseline are marked with * (t-test for automatic and Chi-square for human evaluation). Best scores for each measure are boldfaced.

Does GINGER improve grounding and source attribution over the baseline?

- By operating on information nuggets, **GINGER grounds responses in specific facts and significantly improves source attribution over the baseline**
- Human evaluation shows that the responses generated by baseline and GINGER are comparable with a clear preference towards our method in terms of coherence and conciseness

Method	Automatic evaluation			Human evaluation							
	Entailment	Contradiction	Completeness	Coh	Con	Eng	Fac	Suf	Pref	FQ_rel	FQ_use
Baseline	0.34	0.10	0.25	0.45	0.40	0.50	0.48	**0.52**	0.49	2.58	2.63
GINGER	**0.61***	0.06	0.29	**0.55**	**0.60***	**0.50**	**0.52**	0.48	**0.51**	2.58	2.60

Automatic and human evaluation of responses. Automatic measures target grounding (nugget entailment and contradiction) and completeness. Human evaluation reports the fraction of votes received when compared with the other method for (Coh)erence, (Con)ciseness, (Eng)agingness, (Fac)tuality, (Suf)ficiency, response (Pref)erence, and average scores for follow-up questions (on 3-point Likert scale) in terms of relevance (FQ_rel) and usefulness (FQ_use). Statistically significant differences ($p < 0.05$) with respect to the baseline are marked with ∗ (t-test for automatic and Chi-square for human evaluation). Best scores for each measure are boldfaced.

Is our method capable of generating useful follow-up questions based on facet clusters?

- Follow-up questions generated by our method are on par with the facet-based approach used as our baseline
- **Questions generated by GINGER are guaranteed to be answerable**, which plays a crucial role in dialogue continuity and reliability of the conversational system

Method	Automatic evaluation			Human evaluation							
	Entailment	Contradiction	Completeness	Coh	Con	Eng	Fac	Suf	Pref	FQ_rel	FQ_use
Baseline	0.34	0.10	0.25	0.45	0.40	0.50	0.48	**0.52**	0.49	**2.58**	**2.63**
GINGER	**0.61***	**0.06**	**0.29**	**0.55**	**0.60***	**0.50**	**0.52**	0.48	**0.51**	**2.58**	2.60

Automatic and human evaluation of responses. Automatic measures target grounding (nugget entailment and contradiction) and completeness. Human evaluation reports the fraction of votes received when compared with the other method for (Coh)erence, (Con)ciseness, (Eng)agingness, (Fac)tuality, (Suf)ficiency, response (Pref)erence, and average scores for follow-up questions (on 3-point Likert scale) in terms of relevance (FQ_rel) and usefulness (FQ_use). Statistically significant differences ($p < 0.05$) with respect to the baseline are marked with * (t-test for automatic and Chi-square for human evaluation). Best scores for each measure are boldfaced.

Conclusions

- Answering complex open-ended questions poses multiple challenges and not all of them can be solved by simply applying larger LLMs
- We propose to ensure grounding of the response generated with LLM by operating on information nuggets
- GINGER employs a multi-stage process (clustering, reranking, summarization, fluency enhancement) to generate concise, information-rich, free of redundancy, grounded, completeness-aware conversational responses
- Key advantages of GINGER:
    - maximizing information within response length limits
    - providing source attribution for verifiability
    - guiding users with relevant follow-up questions
    - allowing control over response completeness

GINGER response

What was the US reaction to the Black Lives Matter movement?

Passage 1: On May 26, 2016, Louisiana was the first state to add police officers and firefighters to their state hate crime statute, when Governor John Bel Edwards signed an amendment from the legislature into law. ...

Passage 2: Blue Lives Matter is a pro-police movement in the United States. It was started after the killings of NYPD officers Rafael Ramos and Wenjian Liu in Brooklyn, New York, on December 20, 2014 ...

Passage 3: White Lives Matter is an activist group created in response to Black Lives Matter. In August 2016, the Southern Poverty Law Center added 'White Lives Matter' to its list of hate groups. The group ...

Passage 4: ...

Passage 5: ...

→

Response: The response to the Black Lives Matter movement in the US included the addition of an amendment and the emergence of Blue Lives Matter and All Lives Matter, movements supported by advocates of the police...

Follow-up question: Do you want to learn more about how these arrests have influenced public perception and policy changes regarding racial issues in the US?