Outline

- Background and Introduction
- Language Modeling
- **Open-Domain Dialogue Systems**
  - **Background and Motivation**
  - **Shallow Integration**
  - **Deep Integration**
- Neural Machine Translation
- Conclusion and Outlook

Dialogue Systems

- Dialogue Systems aim to bridge humans and machines with a natural language interface.

JARVIS – Iron Man's Personal Assistant

Baymax – Personal Healthcare Companion

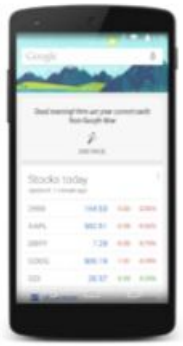
- Humans have long dreamed a machine that understands our languages and responds accordingly.

*Figure [Chen & Gao 17]

Real-world Dialogue Systems

- Dialogue Systems aim to bridge humans and machines with a natural language interface.

Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)

Facebook M & Bot (2015)

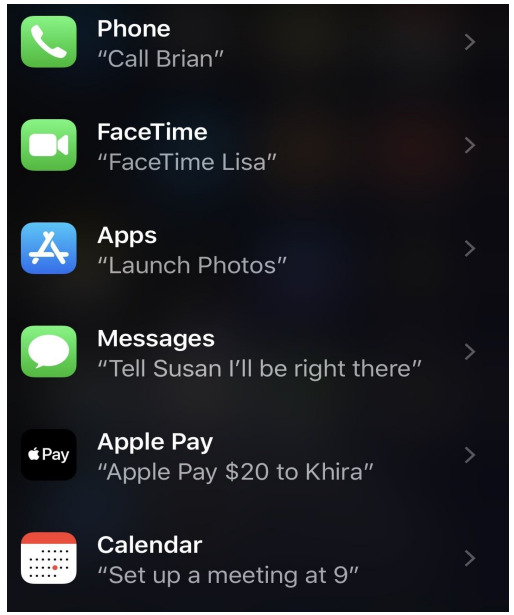
Google Home (2016)

Apple HomePod (2017)

*Figure [Chen & Gao 17]

Categorization of Dialogue Systems

- Dialogue Systems can be categorized into three classes.
  - Task-oriented bot "I need to get this done"
  - Question answering bot "I have a question"
  - Open-domain chit-chat bot "Let's chat for fun"



Apple Siri

IBM Watson won Jeopardy Q&A

Xiaolce

- It is also possible to put them in one chat bot

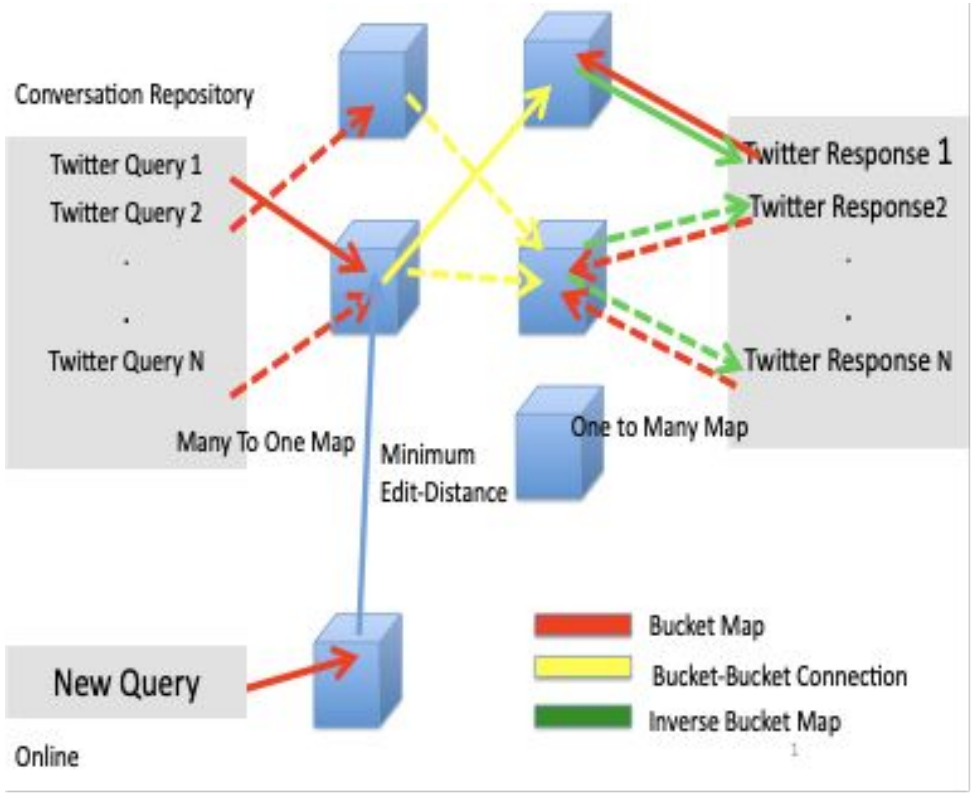
Open-domain Chit-chat Systems

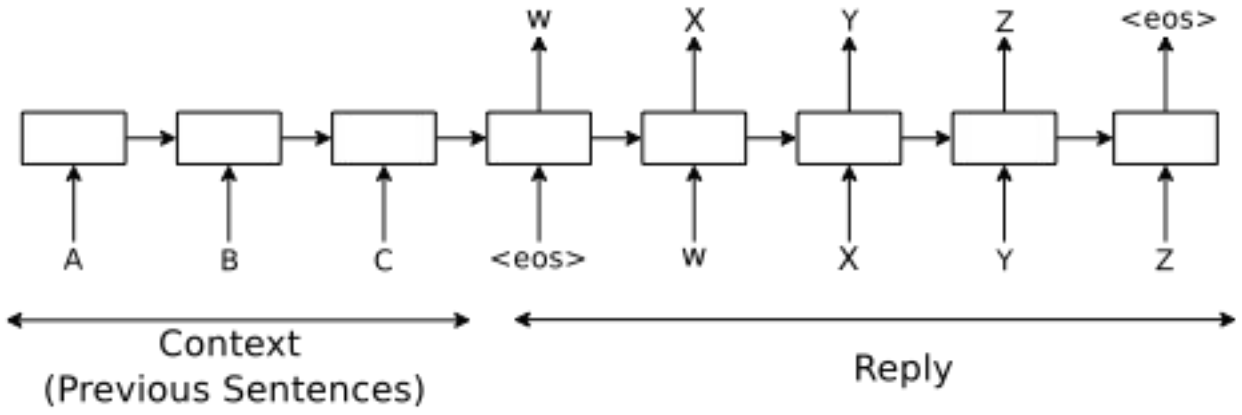
- Dialogue Systems can be categorized into three classes.
  - Task-oriented bot "I need to get this done"
  - Question answering bot "I have a question"
  - Open-domain chit-chat bot "Let's chat for fun"
- Compared to other types, open-domain chit-chat is
  - More open-ended (one-to-many)
  - focused on creating human-like conversations
  - Not restricted in specific domains or tasks

- input: context/query/history
- output: response

Approaches to Open-domain Chit-chat Systems

- Early work in data-driven dialogue response systems
  - retrieval-based [Jafarpour+ 10;Ji+ 14;Hu+ 15]
  - Generation-based [Sordoni+ 15; Vinyals & Le 15; Shang+ 15]

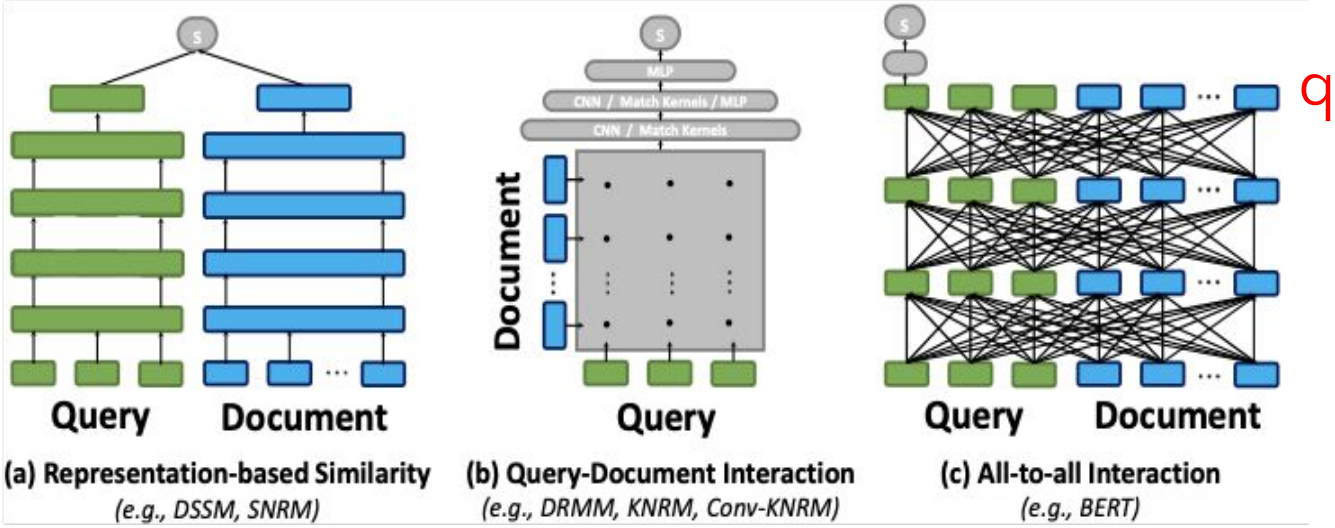




Retrieval-based Dialogue Response Systems

- The **ingredients** of retrieval-based dialogue response systems
  - A (large) database of context-response pairs (or single utterances)
  - A similarity function measuring context-context similarity (e.g, BM25, TFIDF)
  - A relevance function measuring context-response relevance

- Most recent work has been focused on context-response relevance



query-document

classic problem in information retrieval

[Khattab & Zaharia 20]

Pros & Cons of Retrieval-based Systems

- Advantages:
  - fluent
  - informative
  - controllable

written & filtered by humans!

- Disadvantage:
  - This is likely that there is **no** appropriate response in the database

not tailored for input context!

User: How do you like the movie Iron Man?

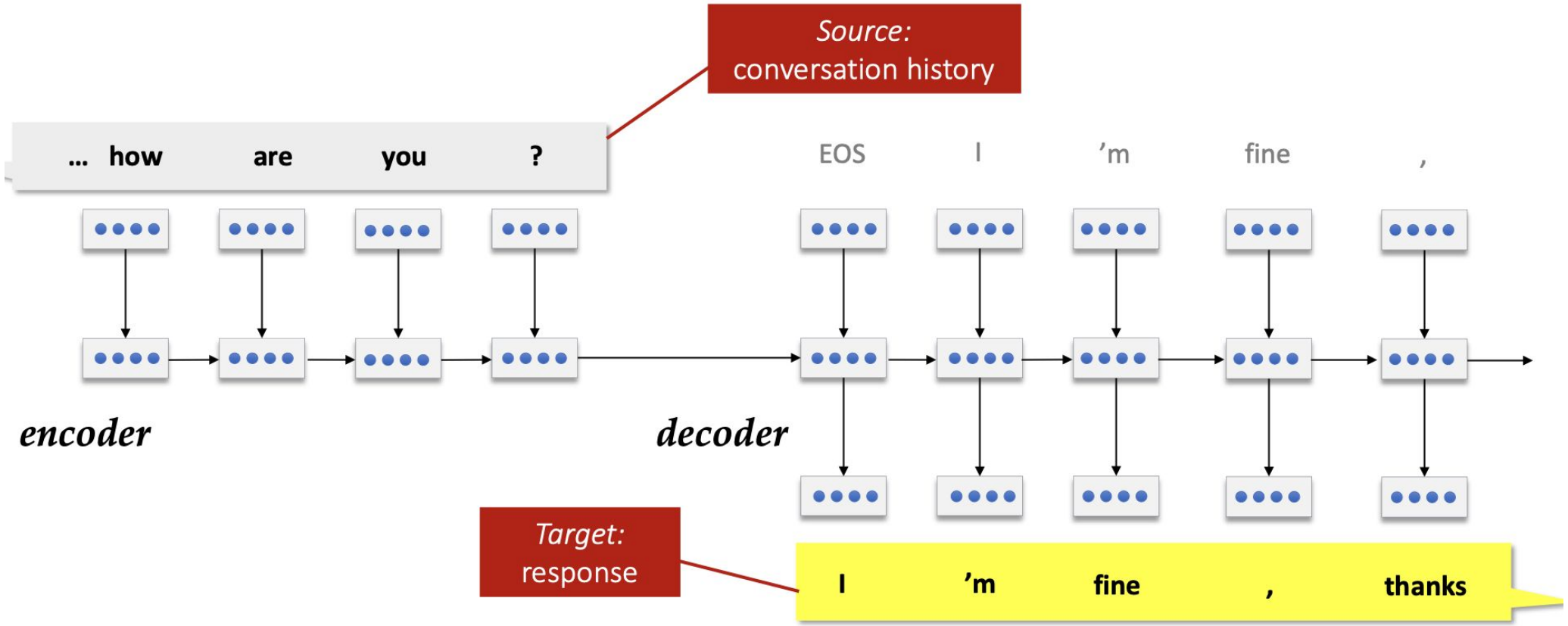
System: Oh, I almost cried when the Batman races to save Rachel.

User: What are you talking about?

* suppose Iron Man is not included the database

Generation-based Dialogue Response Systems

- Generation-based dialogue response systems
  - Seq2Seq (encoder-decoder), similar to neural machine translation
  - RNN/CNN/Transformer etc



[Sordoni+ 15; Vinyals & Le 15; Shang+ 15]

*Figure [Gao+ 18]

Pros & Cons of Generation-based Systems

- Advantages:
  - universal
  - coherent
- Disadvantages:
  - Boring
  - Uninformative
  - Less controllable

it could say anything

Or…just say "**I don't know!**"

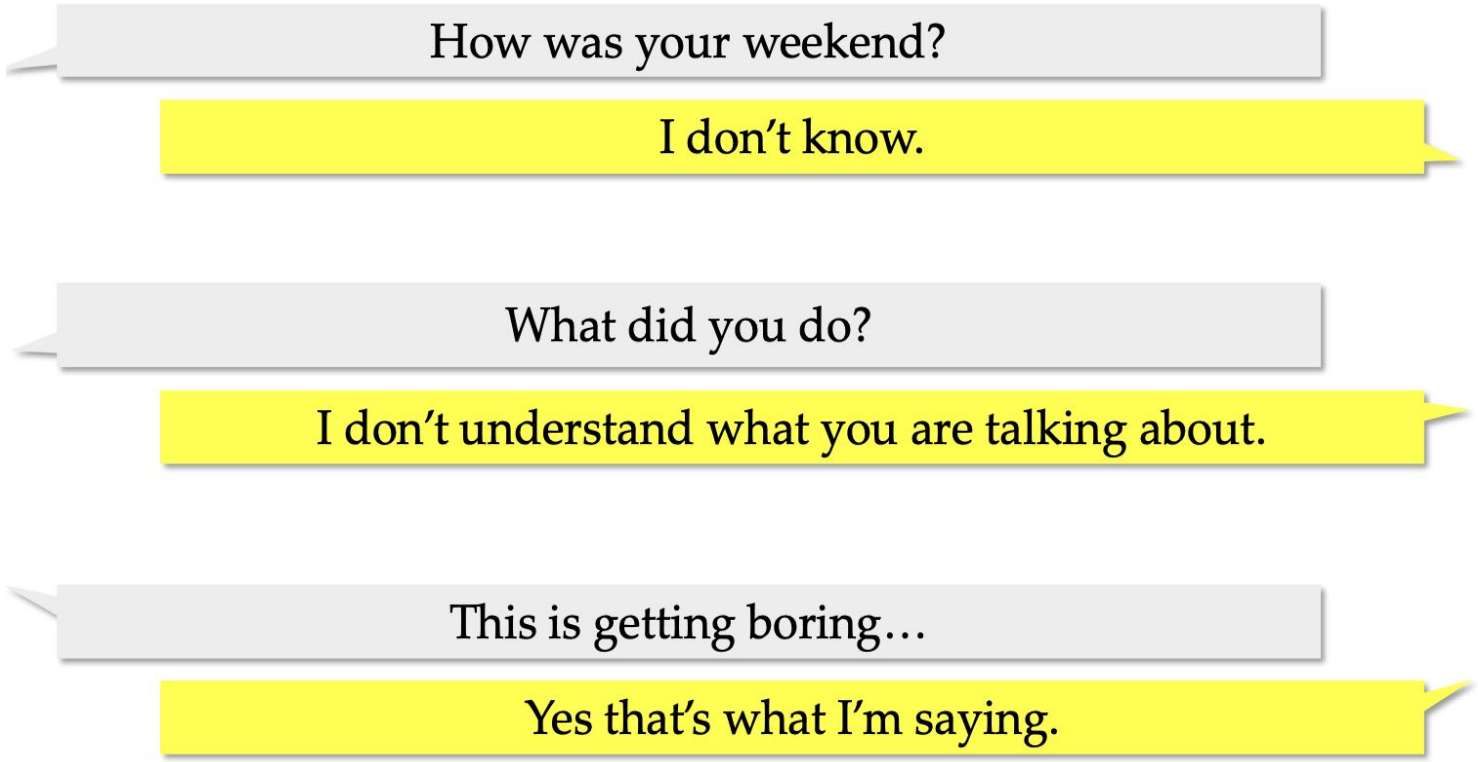
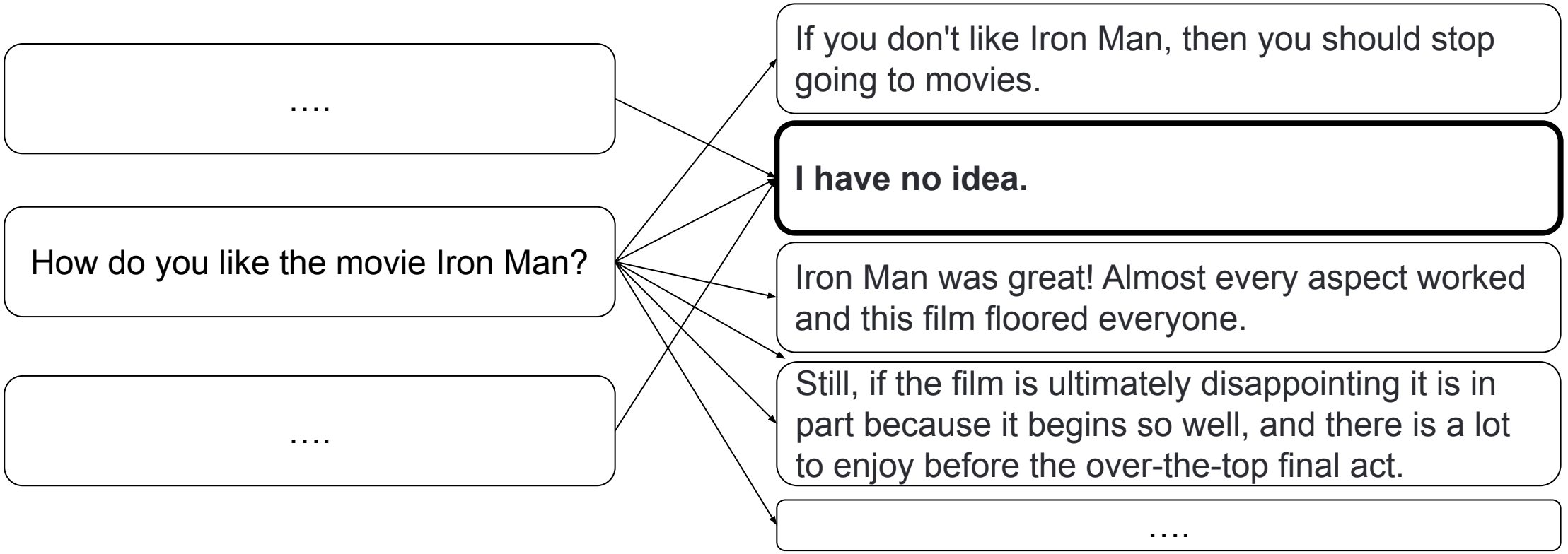


*Figure [Gao+ 18]

Safe Response Problem

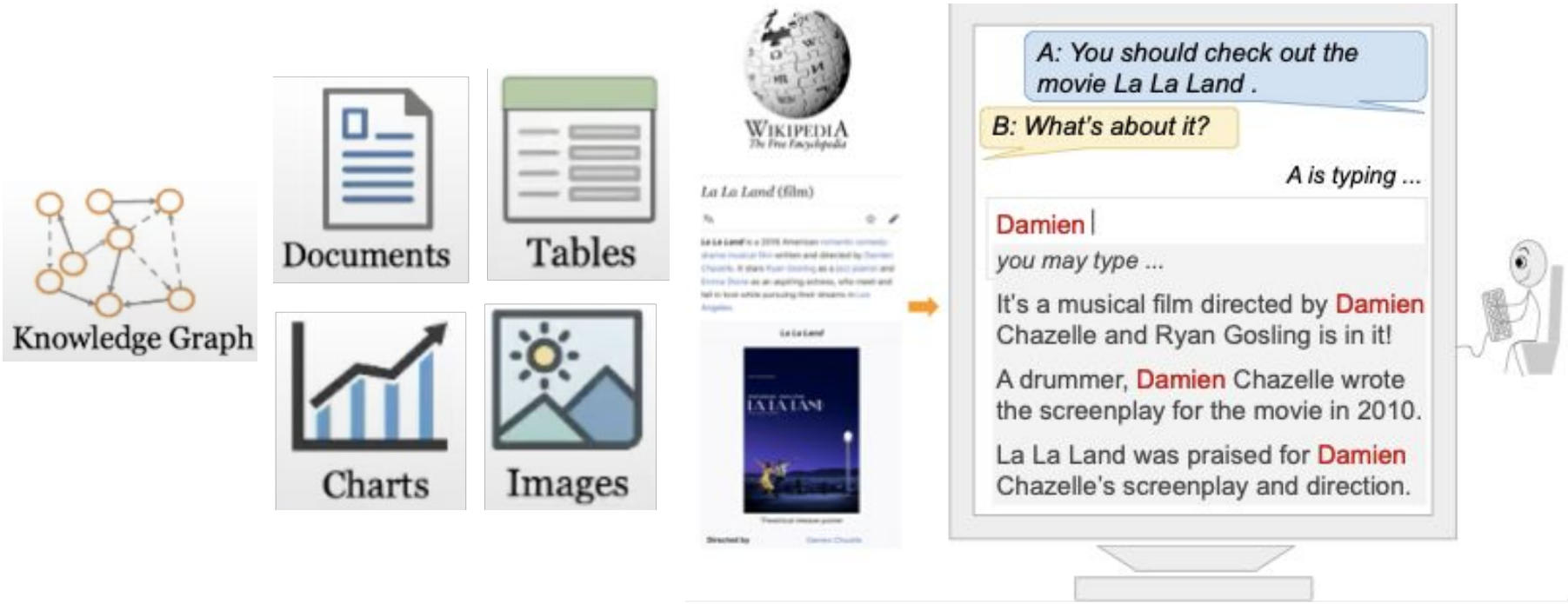
- Safe response problem is one most critical issue in generation-based systems
- Recall the goal of open-domain chit-chat
  - maximize user engagement with informative and enjoyable human-like responses
- Cause: trained models prefer the most common response among others



Remedies for the Safe Response Problem

- One-to-many modeling [Li+ 16; Zhao+ 17; Zhou+ 17; Zhang+ 18; etc]
  - Conditional variational autoencoder, reinforcement Learning, persona, emotion, etc.

- Grounded response generation [Dinan+ 18; Zhou+ 18; Wu+ 21; Komeili+ 22; etc]
  - Grounded on documents, knowledge graphs, images, etc



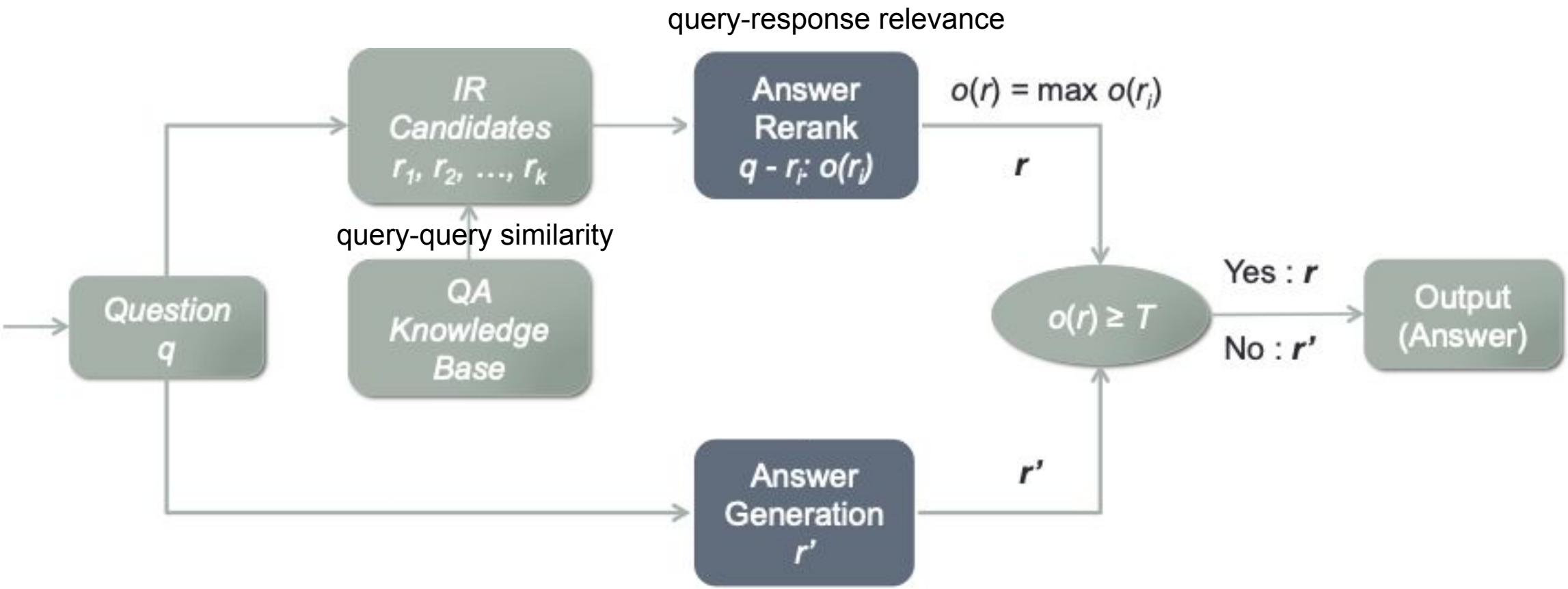
Retrieval vs. Generation

	Retrieval-based Systems	Generation-based Systems
Informativeness	**informative, long**	bland, short
Relevance	good only if similar contexts are in the database	**can generate new responses to unseen contexts**
Controllability	easy to control the database	Blackbox neural models

Retrieval + Generation?

Shallow Integration of Retrieval and Generation

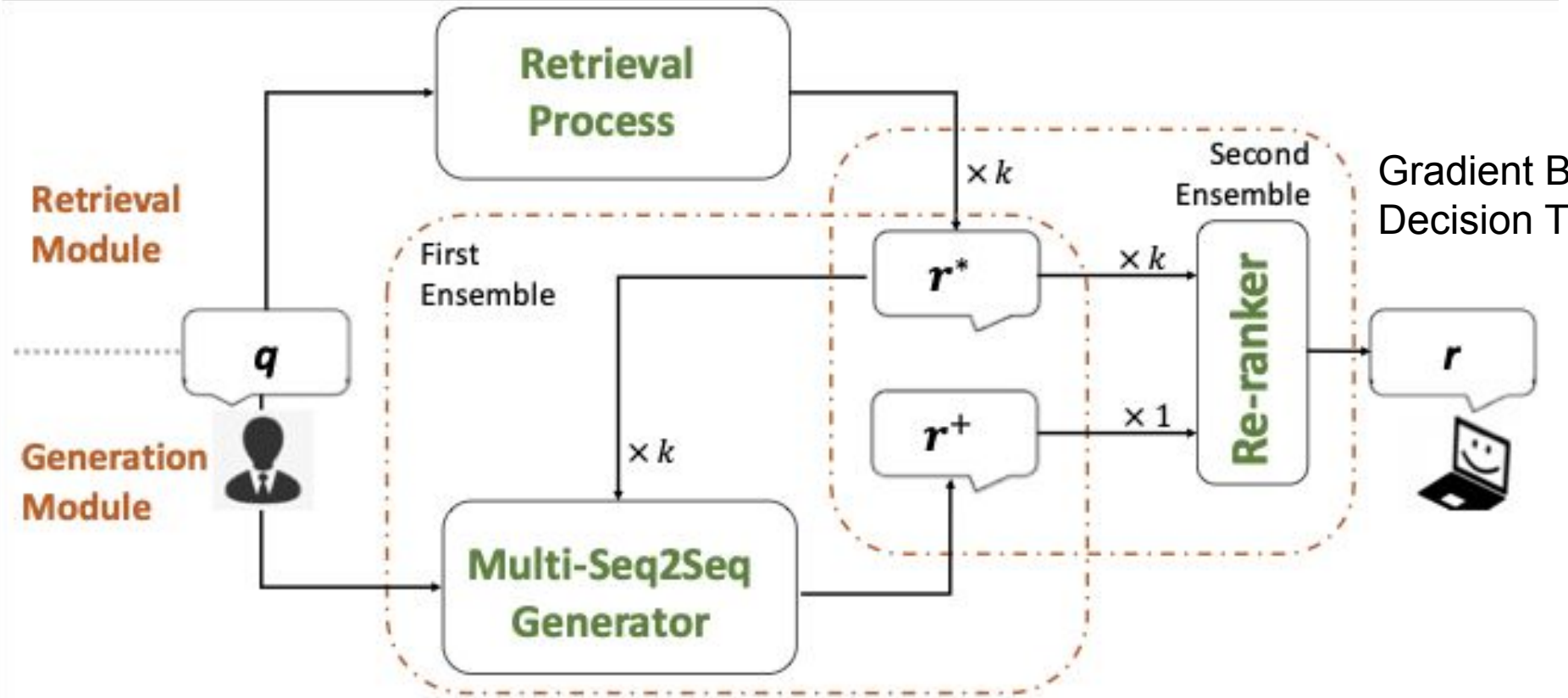
- Switch to generation-based systems when retrieval is "not good"



[Qiu + 17]

Shallow Integration of Retrieval and Generation

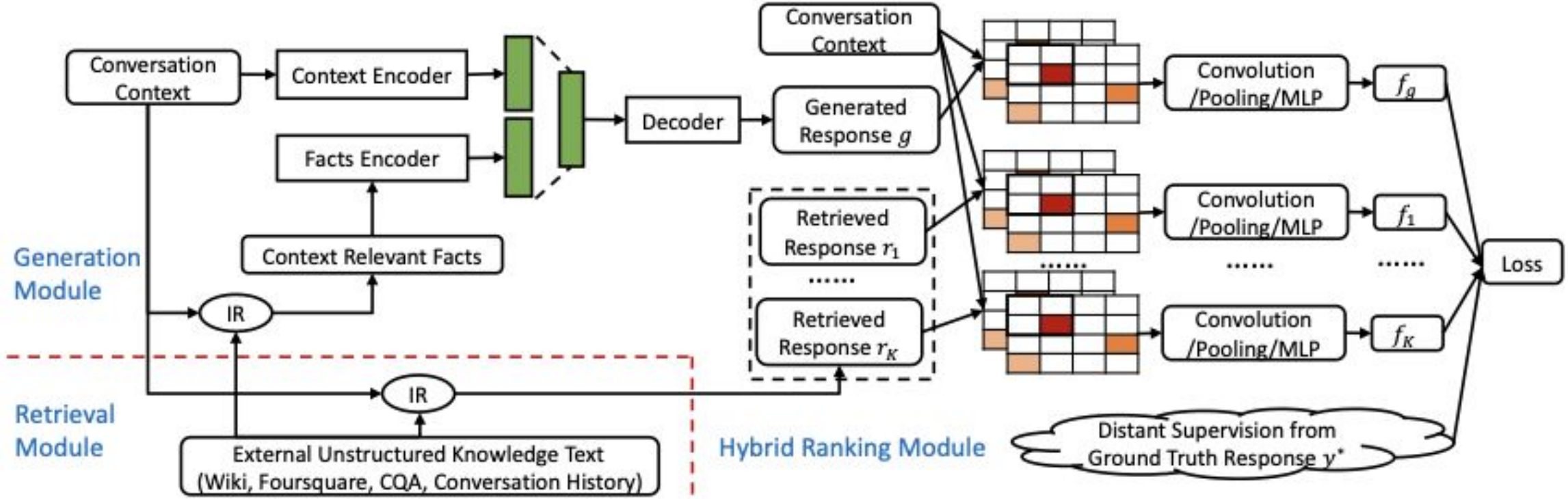
- First Ensemble: Retrieval results are fed into generation-based systems
- Second Ensemble: Rerank all produced responses (generation & retrieval)



[Song + 18]

Shallow Integration of Retrieval and Generation

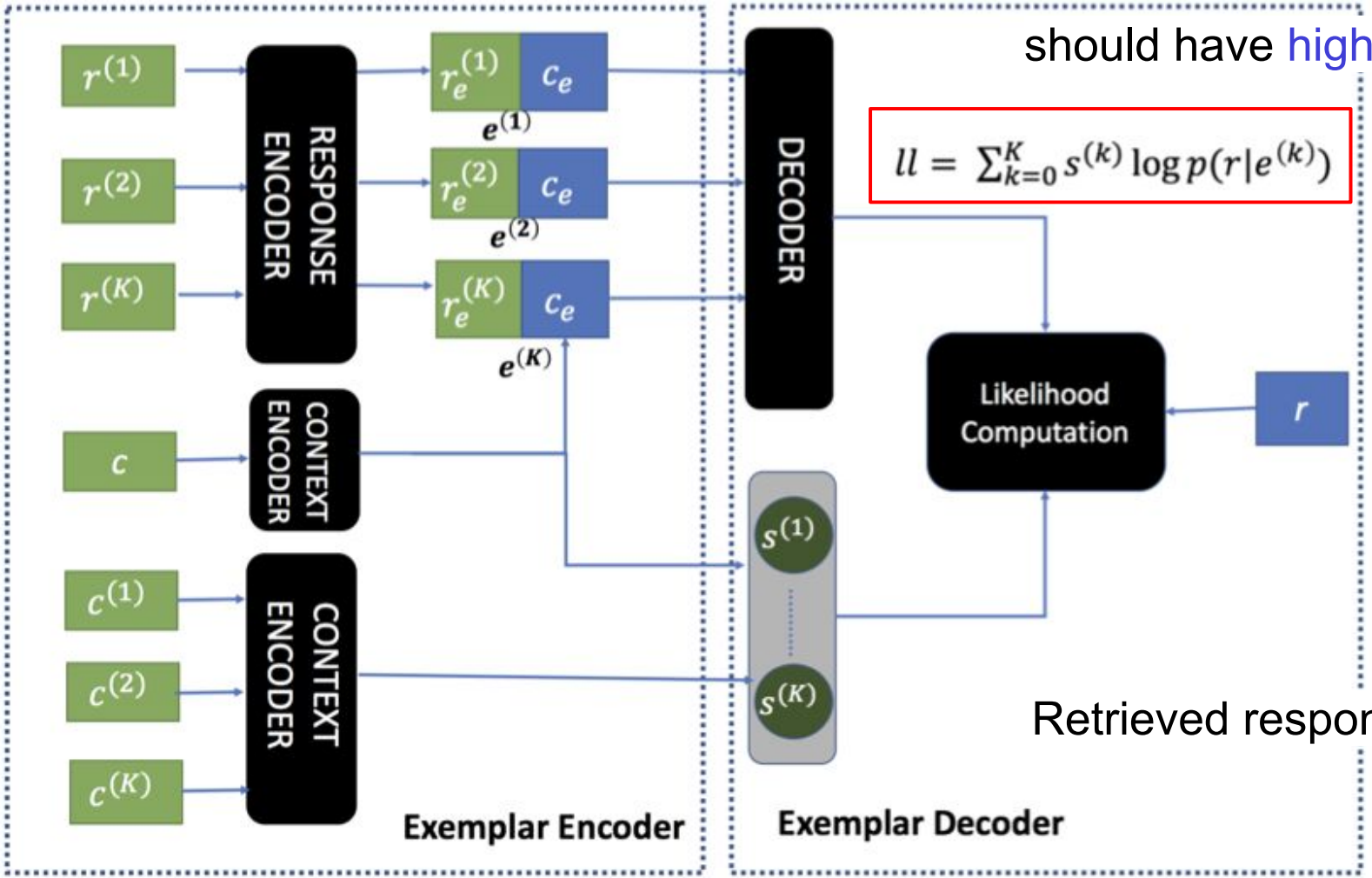
- Improving the Second Ensemble: Rerank all produced responses
  - Model: GBDT => deep neural models
  - Training Data: ground-truth/random negatives => labeled system outputs



[Yang + 19]

Shallow Integration of Retrieval and Generation

- Improving the First Ensemble: retrieval-augmented generation



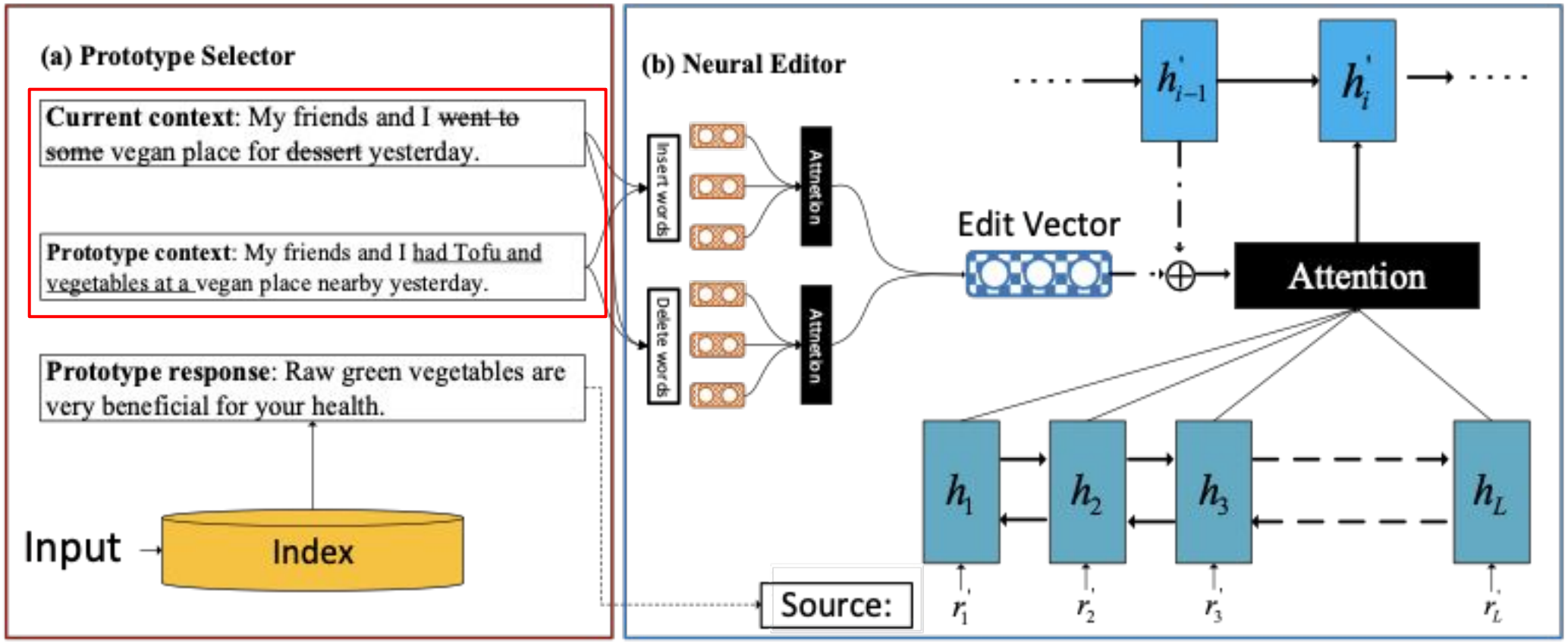
should have higher effect during generation

Retrieved responses with more similar contexts

[Pandey+ 18]

Shallow Integration of Retrieval and Generation

- Improving the First Ensemble: retrieval-augmented generation
  - Differences in **contexts** provide an important signal for differences in **responses**.



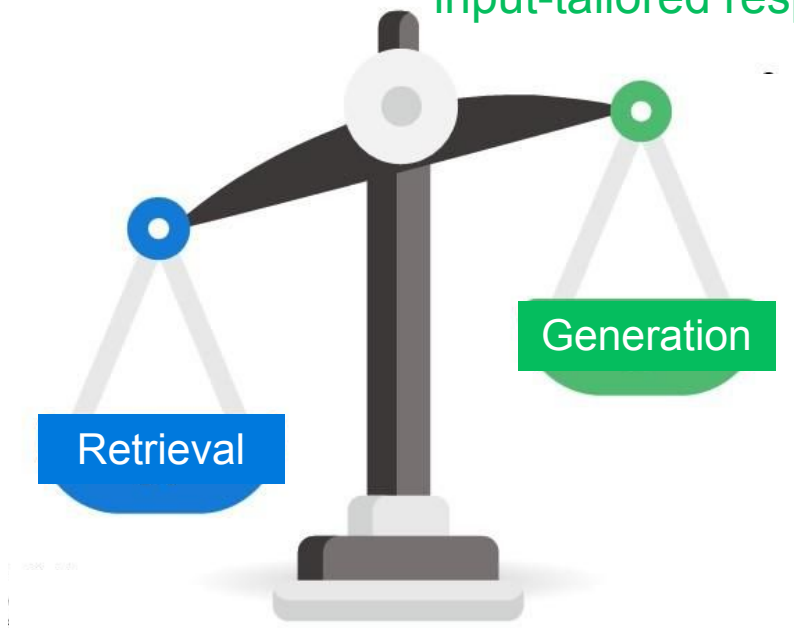
[Wu + 19]

Problems when Integrating Retrieval and Generation

- Collapsing to the ordinary retrieval system

when the retrieval is generally good

lose the ability to make input-tailored responses



overly rely on retrieval
even copy irrelevant content

Filter out irrelevant content from retrieval

The retrieved responses typically contain excessive information, including inappropriate words or entities. It is necessary to filtered out irrelevant content.

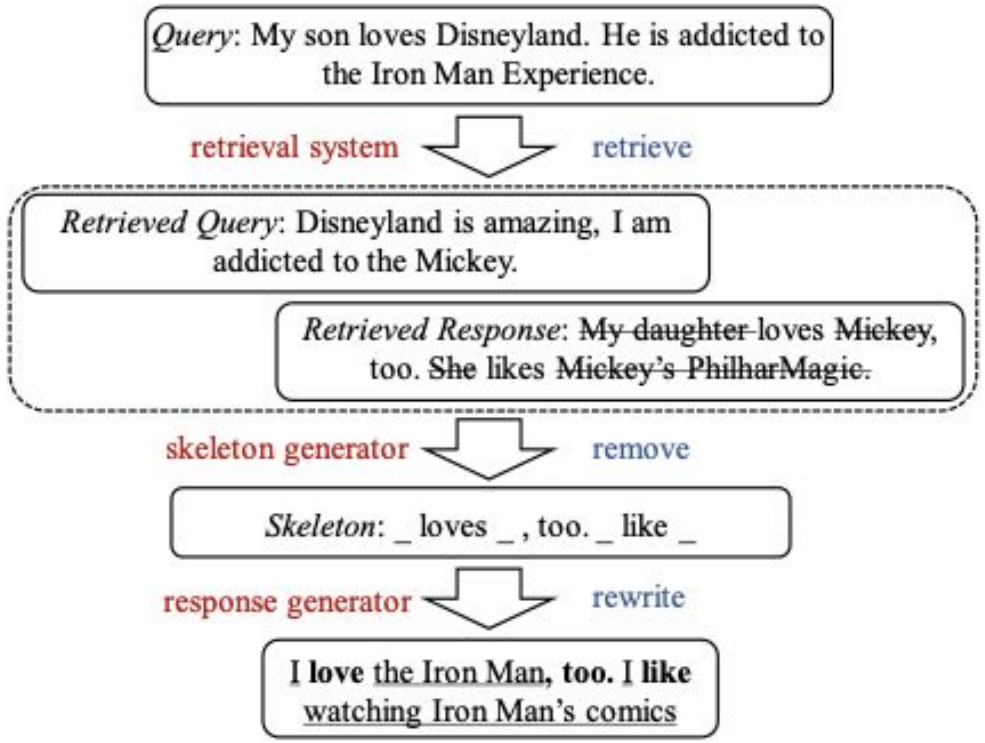
Maintain the generalizability of generation

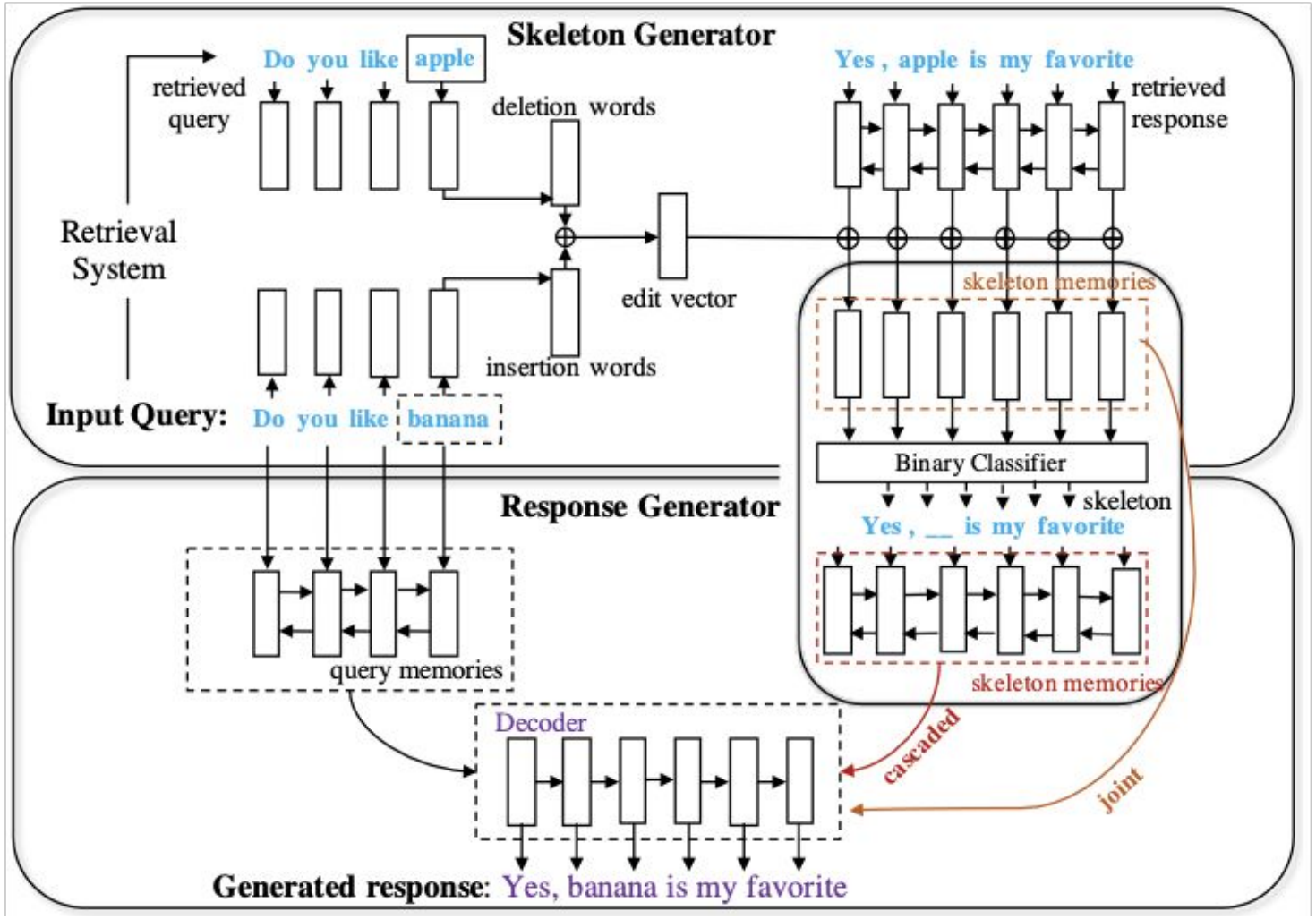
The guidance from retrieval should only specify a response pattern or provide some information, but leave the details to be elaborated by the generation model.

Deep Integration of Retrieval and Generation

- Retrieve-Remove-Rewrite
  - extracting **response skeleton**

explicitly control the information inflow





[Cai + 19]

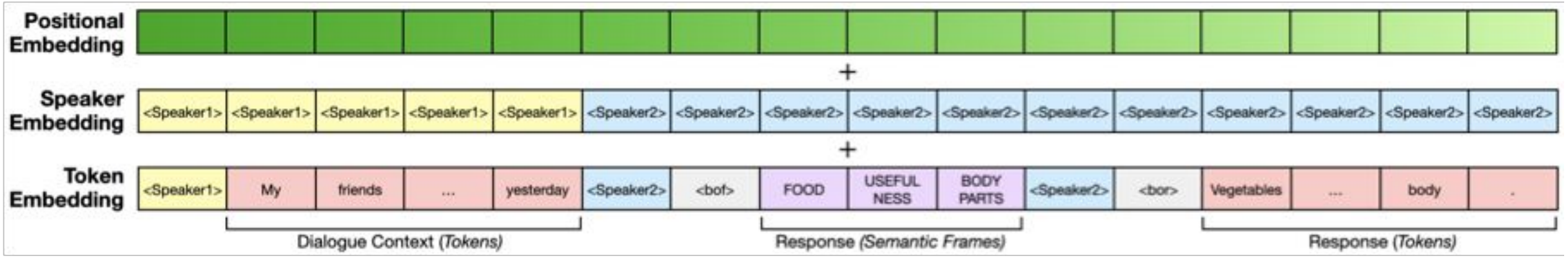
Deep Integration of Retrieval and Generation

- Retrieve-Abstract-Follow
  - extracting **semantic structure**

preserve the semantic structure

avoid over-reliant on copying (inappropriate) words

Context	My friends and I have started eating vegan food since yesterday.
Exemplar	Eggs are very beneficial for your body.
Frames	FOOD USEFULNESS BODY-PARTS
Responses	Vegan food can be good for your health. Vegetables can do wonders for your body Vegan food is very healthy.
Exemplar	I want to drink milk as well.
Frames	DESIRING INGESTION FOOD
Responses	You want to eat some vegan food? We eat a lot of vegetables. It's delicious. We like to eat organic food.



[Gupta + 21]

Deep Integration of Retrieval and Generation

Model	Dist-2	Dist-3	MaUdE	Coherent	Fluent	Consistent	Interesting
Retrieval	0.294	0.526	0.921	2.41	2.61	2.48	2.32
GPT2-Gen	0.249	0.494	0.905	2.42	2.55	2.41*	2.18*
LSTM-Tokens	0.182	0.380	0.890	2.04*	2.10*	2.11*	1.89*
LSTM-Frames	0.185	0.392	0.901	2.36*	2.30*	2.33*	1.97*
GPT2-Tokens	0.254	0.513	**0.927**	2.19*	2.47*	2.29*	2.11*
EDGE (Ours)	**0.278**	**0.571**	0.922	**2.52**	**2.63**	**2.56**	**2.39**
Human	0.385	0.720	0.911	2.76	2.69	2.78	2.44

Context	*Human1*: they sell everything. *Human2*: well, i want chinese food.	*Human1*: actually i have a passion for chinese literature. *Human2*: you do?
Retrieved	well, what do you want to eat?	yes, reading is my hobby.
Frames	WHAT DESIRING INGESTION ?	YES LINGUISTIC-MEANING
GPT2-Gen	it's a good idea.	yes. i'm passionate.
LSTM-Tokens	well, what's the you do?	yes, i do.
LSTM-Frames	i hope so.	yes, i did.
GPT2-Tokens	i'm not sure what to get.	what are you interested in?
EDGE (Ours)	you want to eat something chinese?	yes. i studied chinese literature at university.

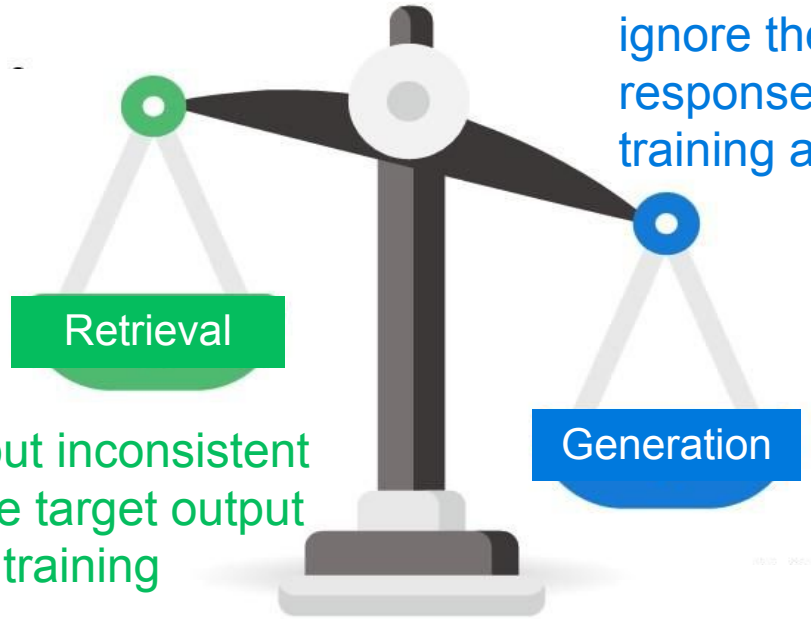
[Gupta + 21]

Problems when Integrating Retrieval and Generation

- Collapsing to the ordinary generation system

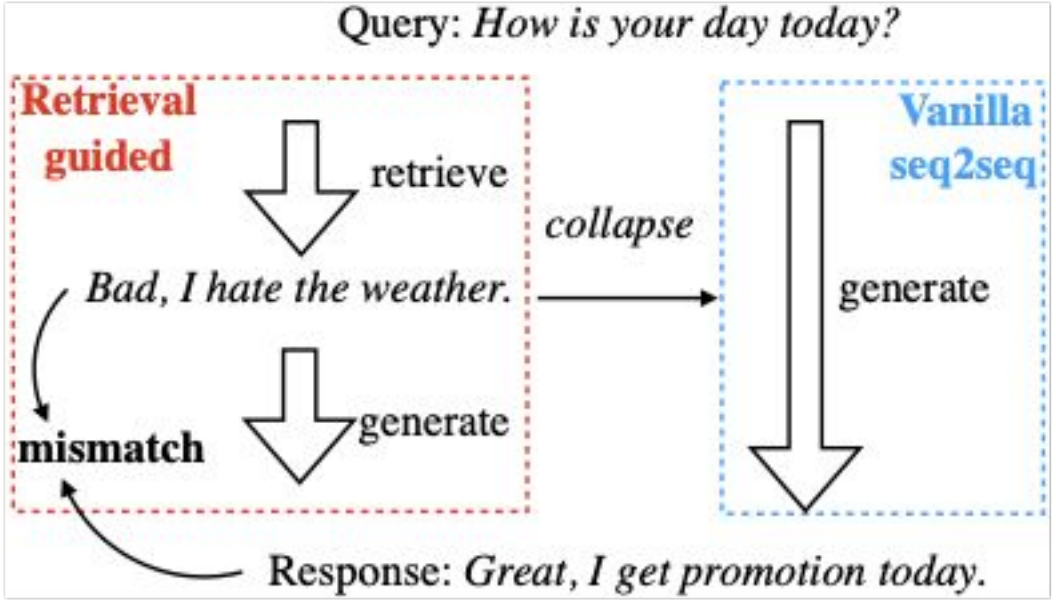
inconsistent context-retrieval-response triples for training

context-relevant ≠ response-relevant



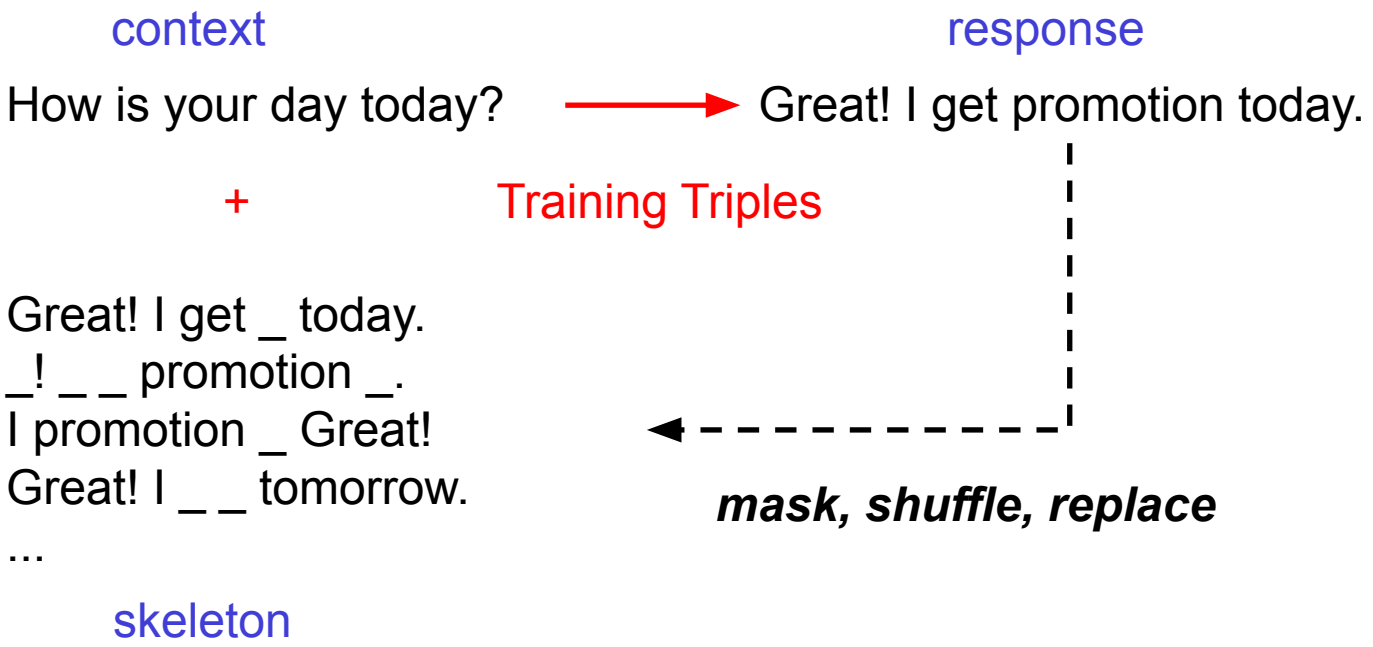
ignore the retrieved responses during training and testing

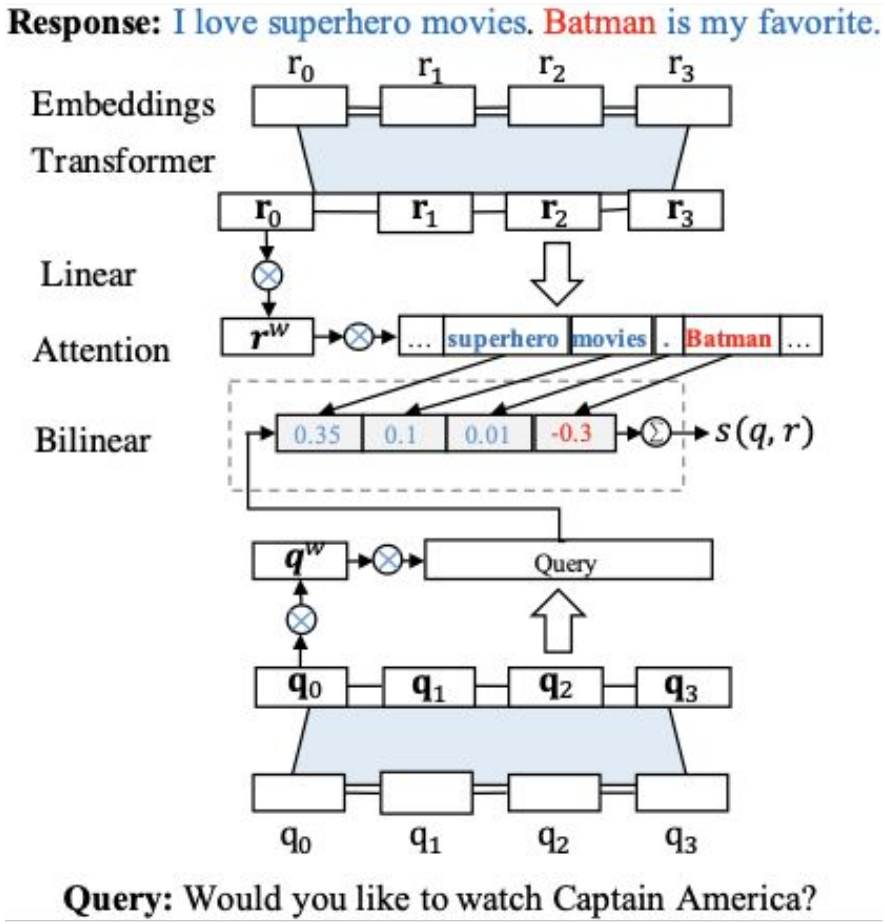
good but inconsistent with the target output during training



Deep Integration of Retrieval and Generation

- Response-consistent skeletons generated automatically from the target response
- Accurate skeleton extraction with distant supervision from semantic matching





[Cai + 19]

Deep Integration of Retrieval and Generation

Tencent AI Lab

- Improve the best of two worlds:
  - Higher informativeness than vanilla retrieval
  - Higher relevance than vanilla generation

Models	Informativeness	Relevance	Fluency
Retrieval	2.65 (0.90)†	2.58 (0.86)	2.96 (0.72)
Seq2Seq	2.01 (0.65)	2.58 (0.53)	2.71 (0.43)
Seq2Seq-MMI	2.47 (0.70)	2.79 (0.67)	2.99 (0.61)
RetrieveNRefine^{++}	2.30 (0.79)	2.62 (0.63)	2.82 (0.51)
EditVec	2.29 (0.61)	2.62 (0.60)	2.83 (0.47)
Skeleton-Lex	2.45 (0.61)	2.80 (0.56)	2.99 (0.46)
Ours	**2.69** (0.87)	**3.11** (0.55)	**3.20** (0.55)

[Cai + 19]

Deep Integration of Retrieval and Generation

- Model response-posterior distribution

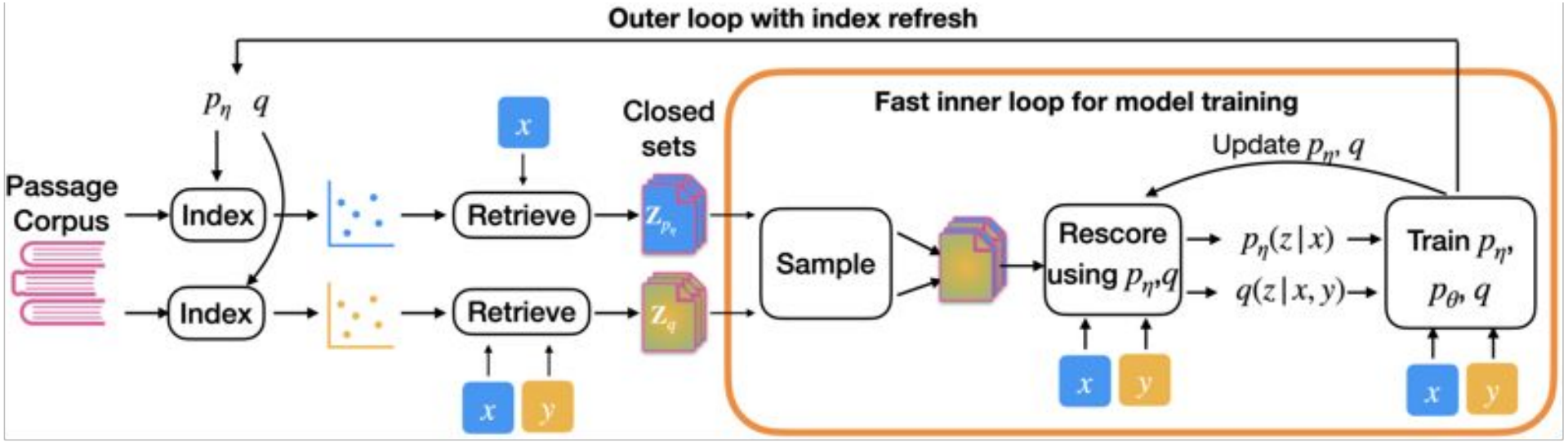
$$P(y|x) = \sum_{z \in \text{top-k}(P_\eta(.|x))} P_\eta(z|x) P_\theta(y|x,z)$$

retriever generator

$$\log P(y|x) \geq \mathbb{E}_{z \sim Q(.|x,y)}[\log P_\theta(y|x,z)] - D_{\mathrm{KL}}(Q|P_\eta)$$

response-posterior

- differentiate response-relevant from other context-relevant retrieval
- encourage the retriever to trust response-relevant



[Paranjape + 21]

Takeaways

Tencent AI Lab

- Retrieval helps generation in open-domain dialogues
  - promote **informativeness** and **relevance**
  - provide **explainability** and **controllability**
- but… should be used with caution for the following problems
  - Information overflow (overly rely on retrieval)
  - Inconsistent context-retrieval-response training triples (ignore retrieval)