Doug Downey, Allen Institute for AI

Chasing high-precision NLP at discount prices: Lessons for accelerating science

Motivation

Scientists remain beset by information overload

Recent LLM breakthroughs present tantalizing possibilities for helping scientists to **discover**, **understand**, and **synthesize** literature

Motivation

Scientists remain beset by information overload

Recent LLM breakthroughs present tantalizing possibilities for helping scientists to **discover**, **understand**, and **synthesize** literature

So far, the hype has outpaced the reality

Two major barriers: **inaccuracy** and **cost**

Example: Hallucination

User: What’s unique k-sat and what are some good recent papers on it?

GPT-4:

Unique k-SAT is a variation of the Boolean satisfiability problem (k-SAT) where **the solution is unique**. The problem **involves determining the truth assignment for variables** in a Boolean formula, with the constraint that each clause consists of at most 'k' literals and **exactly one assignment will make the entire formula true**.

Example: Hallucination

User: What’s unique k-sat and what are some good recent papers on it?

GPT-4:

Unique k-SAT is a variation of the Boolean satisfiability problem (k-SAT) where **the solution is unique**. The problem **involves determining the truth assignment for variables** in a Boolean formula, with the constraint that each clause consists of at most 'k' literals and **exactly one assignment will make the entire formula true**.

Subtly and importantly wrong

Example: Hallucination

User: What's unique k-sat and what are some good recent papers on it?

GPT-4 (corrected):

Unique k-SAT is a variation of the Boolean satisfiability problem (k-SAT) where **the solution is unique if it exists**. The problem **involves determining** ~~the truth~~ **whether there exists a formula-satisfying assignment for variables** in a Boolean formula, with the constraint that each clause consists of at most 'k' literals and **exactly one or zero assignment will make the entire formula true**.

Subtly and importantly wrong

Example: Hallucination

(GPT-4 cont.)

In the last few years, there have been several studies carried out to understand unique k-SAT problems. Here are some notable papers:

1. "On the Fine-Grained Complexity of One-Dimensional Dynamic Programming" by Marvin Künnemann, Ramamohan Paturi, and Stefan Schneider (2020). This paper refines the current understanding of the fine-grained complexity of unique 3-SAT.

2. "Unique SAT and the Isolation Lemma: A Survey" by Gabriele Röger (2019). This paper provides a comprehensive survey of the unique-SAT problem and its significance in computer science and artificial intelligence.

3. "A Study of Unique and Almost-Unique-SAT" by Shraddha Rathi and Lane A. Hemaspaandra (2018). This paper examined the complexity of unique-SAT.

4. "The Complexity of the Unique Coverage Problem" by Lukasz Kaiser (2005). This paper investigates the complexity of the unique coverage problem, a variant of unique k-SAT where each variable appears in exactly one clause.

Example: Hallucination

(GPT-4 cont.)

In the last few years, there have been several studies carried out to understand unique k-SAT problems. Here are some notable papers:

1. "On the Fine-Grained Complexity of One-Dimensional Dynamic Programming" by Marvin Künnemann, Ramamohan Paturi, and Stefan Schneider (2020). This paper refines the current understanding of the fine-grained complexity of unique 3-SAT.
2. "Unique SAT and the Isolation Lemma: A Survey" by Gabriele Röger (2019). This paper provides a comprehensive survey of the unique-SAT problem and its significance in computer science and artificial intelligence.
3. "A Study of Unique and Almost-Unique-SAT" by Shraddha Rathi and ed the complexity of unique-SAT.
4. "The Complexity of the Unique Coverage Problem" by Lukasz Kaiser (2005). This paper investigates the complexity of the unique coverage problem, a variant of unique k-SAT where each variable appears in exactly one clause.

~Real papers, but not relevant

Example: Hallucination

(GPT-4 cont.)

In the last few years, there have been several studies carried out to understand unique k-SAT problems. Here are some notable papers:

1. "On the Fine-Grained Complexity of One-Dimensional Dynamic Programming" by Marvin Künnemann, Ramamohan Paturi, and Stefan Schneider (2020). This paper refines the current understanding of the fine-grained complexity of unique 3-SAT.

2. "Unique SAT and the Isolation Lemma: A Survey" by Gabriele Röger (2019). This paper provides a comprehensive survey of the unique-SAT problem and its significance in computer science and artificial intelligence.

3. "A Study of Unique and Almost-Unique-SAT" by Shraddha Rathi and ed the complexity of unique-SAT.

4. "The Complexity of the Unique Coverage Problem" by Lukasz Kaiser (2005). This paper investigates the complexity of the unique coverage problem, a variant of unique k-SAT where each variable appears in exactly one clause.

~Real papers, but not relevant

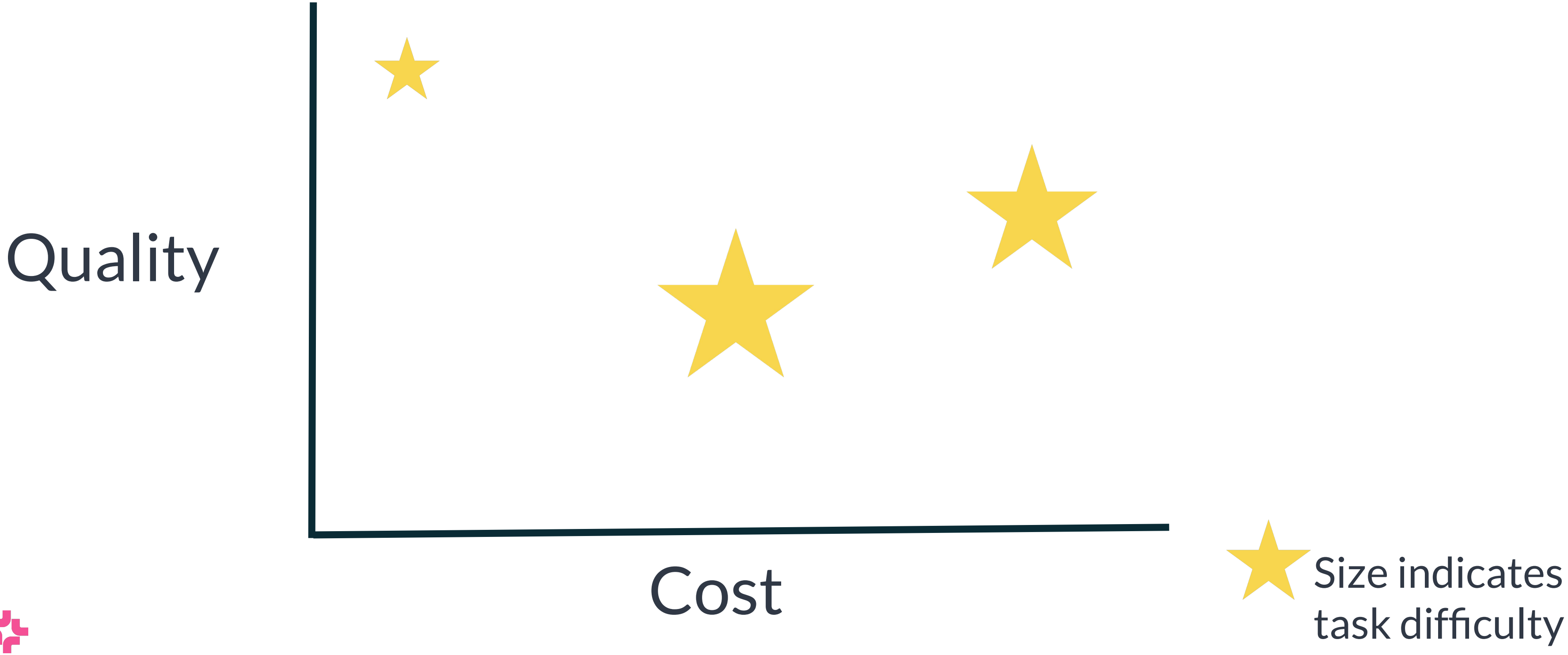
Made-up papers!

RAG to the rescue?

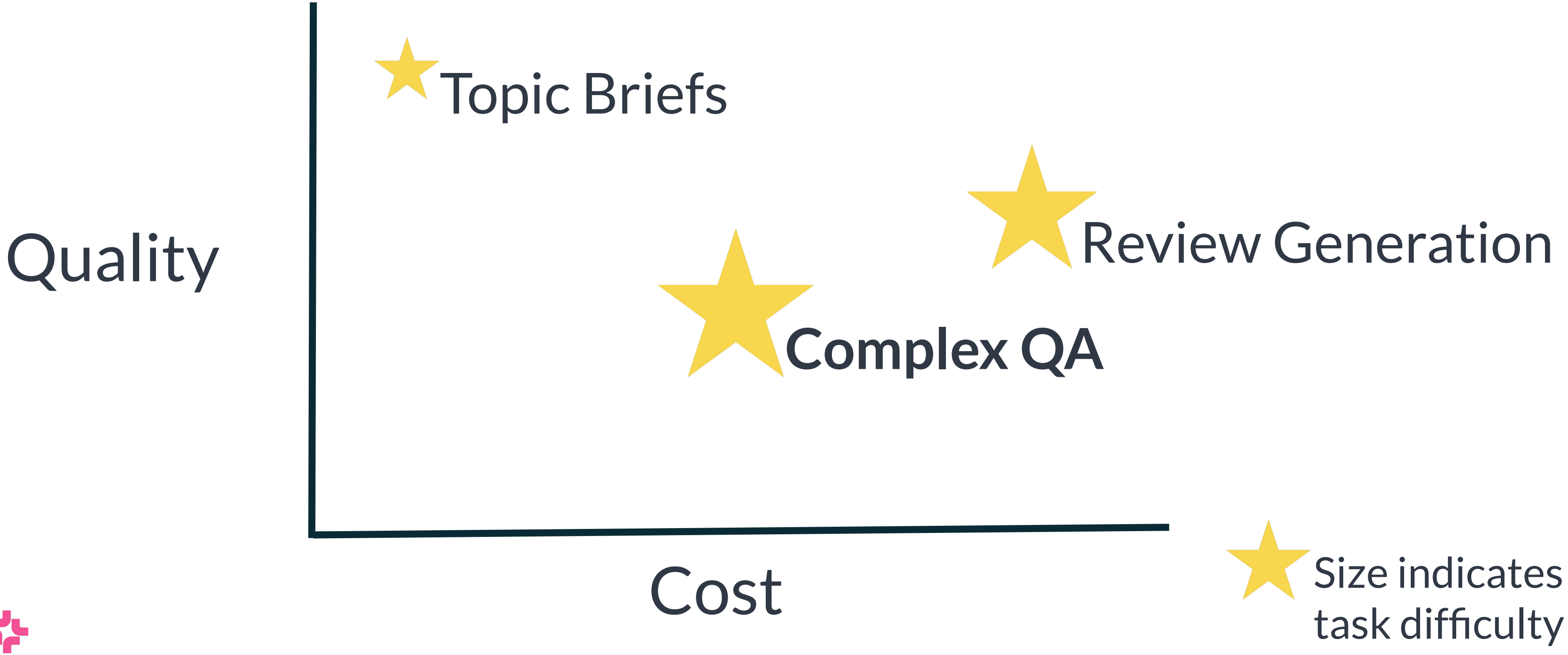
Retrieval Augmented Generation (RAG) grounds responses in literature

Still has significant accuracy limitations, as we will see, and is expensive

Three Tasks, Three LLM-powered Workflows



Three Tasks, Three LLM-powered Workflows



Quality

Topic Briefs

Review Generation

Complex QA

Cost

Size indicates task difficulty

Outline

Application	Scope	Cost	Accuracy
Complex QA	🌎 🌎 🌎 🌎	$ $ $	★ ★
Generating Reviews	🌎 🌎	$ $ $ $ $	★ ★ ★
Scientific Topic Pages	🌎	$	★ ★ ★ ★
If time: PDF Parsing	🌎	$	★ ★ ★ ★ ⯨

Complex Question Answering

Free-form questions that require **multiple documents** to answer, e.g.:

- How is diversity typically evaluated in recommendation systems?
- What data preprocessing steps are most important for point cloud datasets before performing surface reconstruction?
- Is there any evidence that large language models can be effectively applied to robot planning tasks?
- What are some HCI systems papers that conduct an observational study as the formative study instead of just interviews?
- Etc., etc.

Sergey Feldman

Amanpreet Singh

Joseph Chee Chang

Demo

System Architecture

Retrieval

Dense (GIST embeddings) + Sparse (BM25, SparseEmbed) Vespa index

Filtering [expensive!!]

Select relevant snippets with LLM (Claude 3.5 Sonnet today), then sub-select exact quotes

Clustering

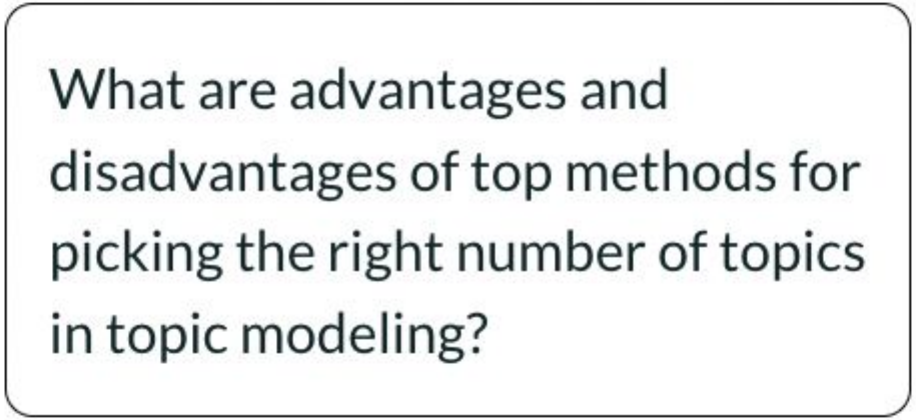
Identify organizing themes, assign snippets to themes

Assembly

Write the response one theme at a time

See also e.g. (Slobodkin et al., 2024)

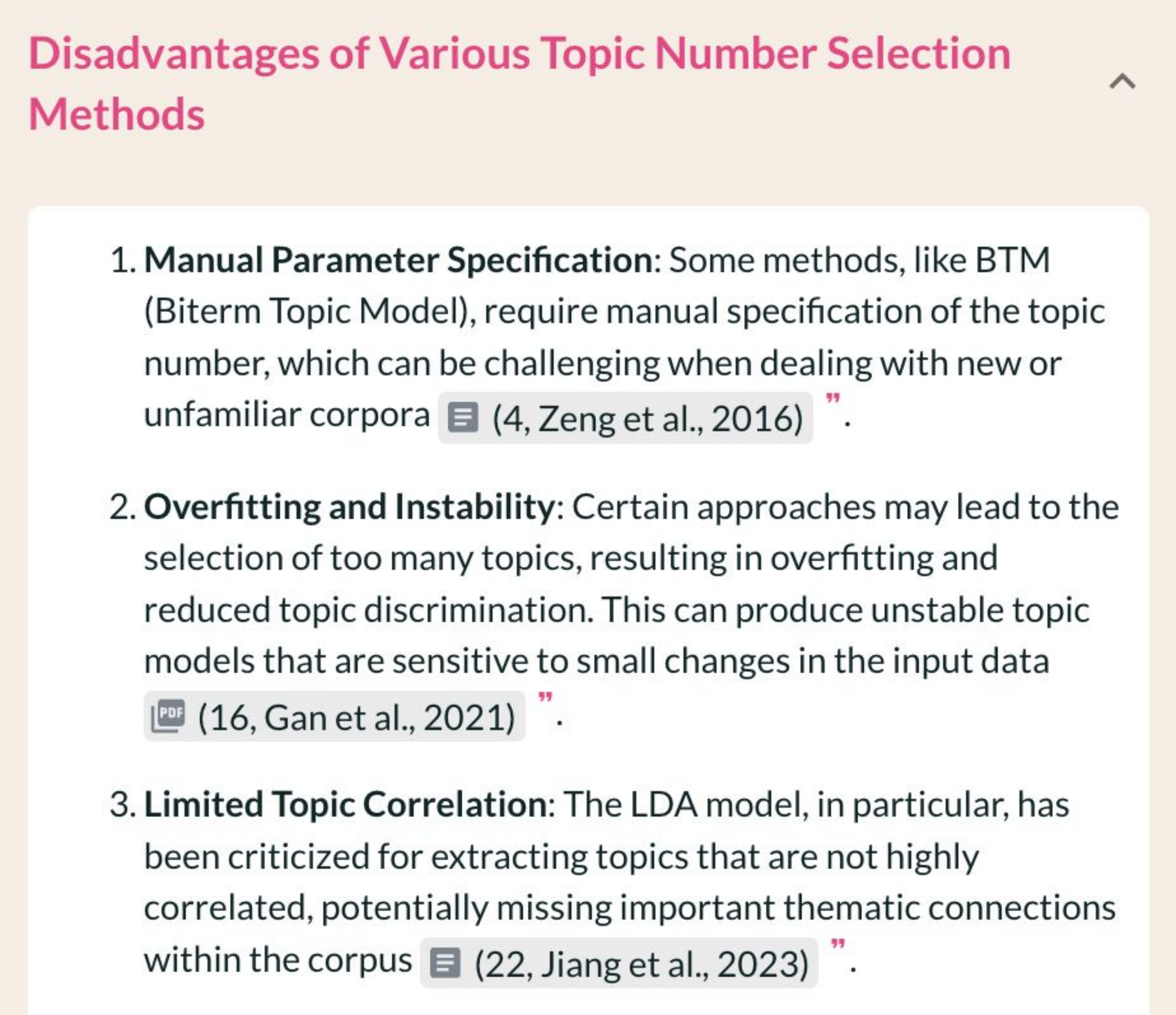
Open Challenges



What are advantages and disadvantages of top methods for picking the right number of topics in topic modeling?

Open Ch

What ar
disadva
picking
in topic

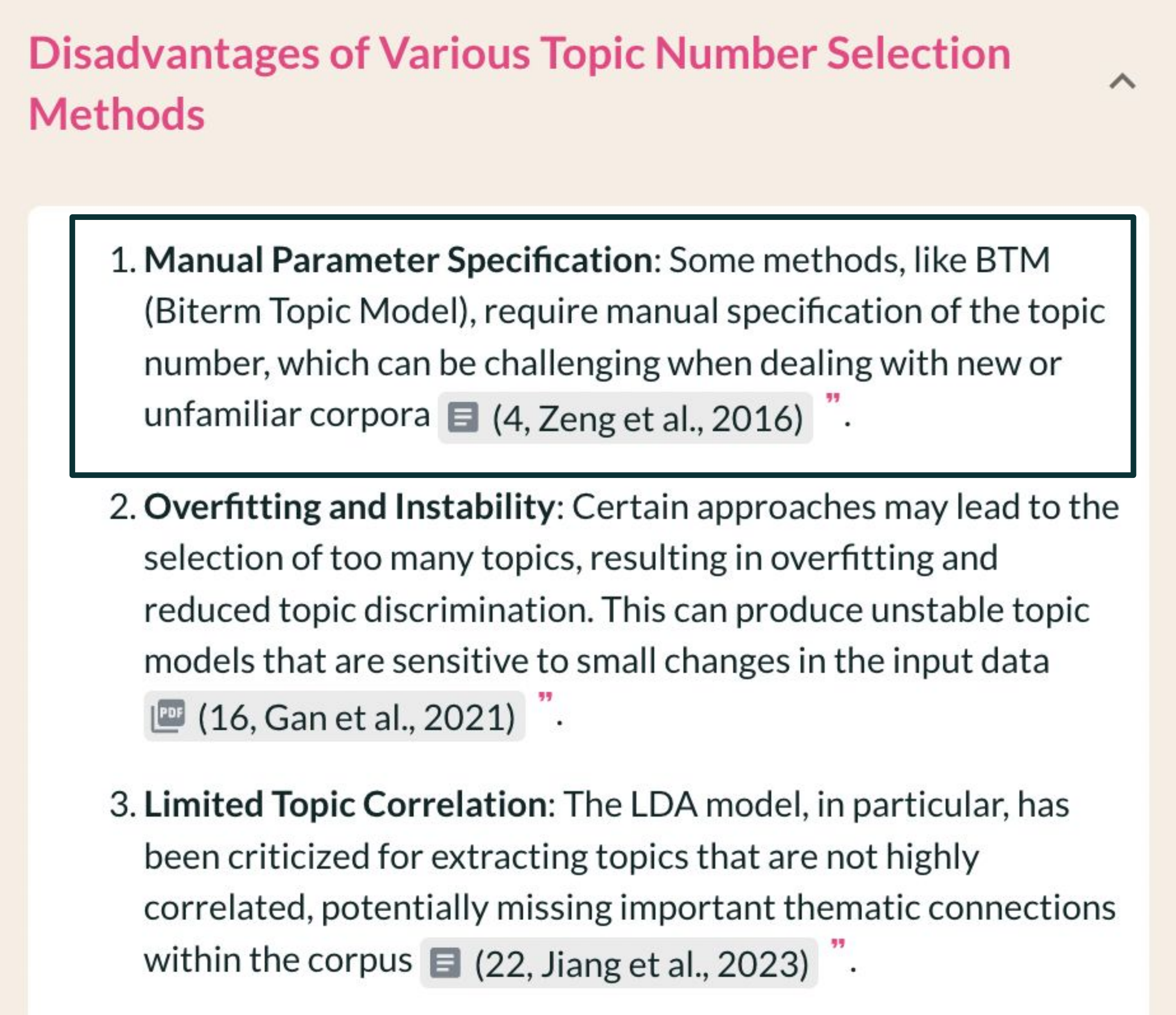


Disadvantages of Various Topic Number Selection Methods

1. **Manual Parameter Specification**: Some methods, like BTM (Biterm Topic Model), require manual specification of the topic number, which can be challenging when dealing with new or unfamiliar corpora (4, Zeng et al., 2016) ”.
2. **Overfitting and Instability**: Certain approaches may lead to the selection of too many topics, resulting in overfitting and reduced topic discrimination. This can produce unstable topic models that are sensitive to small changes in the input data (16, Gan et al., 2021) ”.
3. **Limited Topic Correlation**: The LDA model, in particular, has been criticized for extracting topics that are not highly correlated, potentially missing important thematic connections within the corpus (22, Jiang et al., 2023) ”.

Open Cha

What ar
disadva
picking
in topic

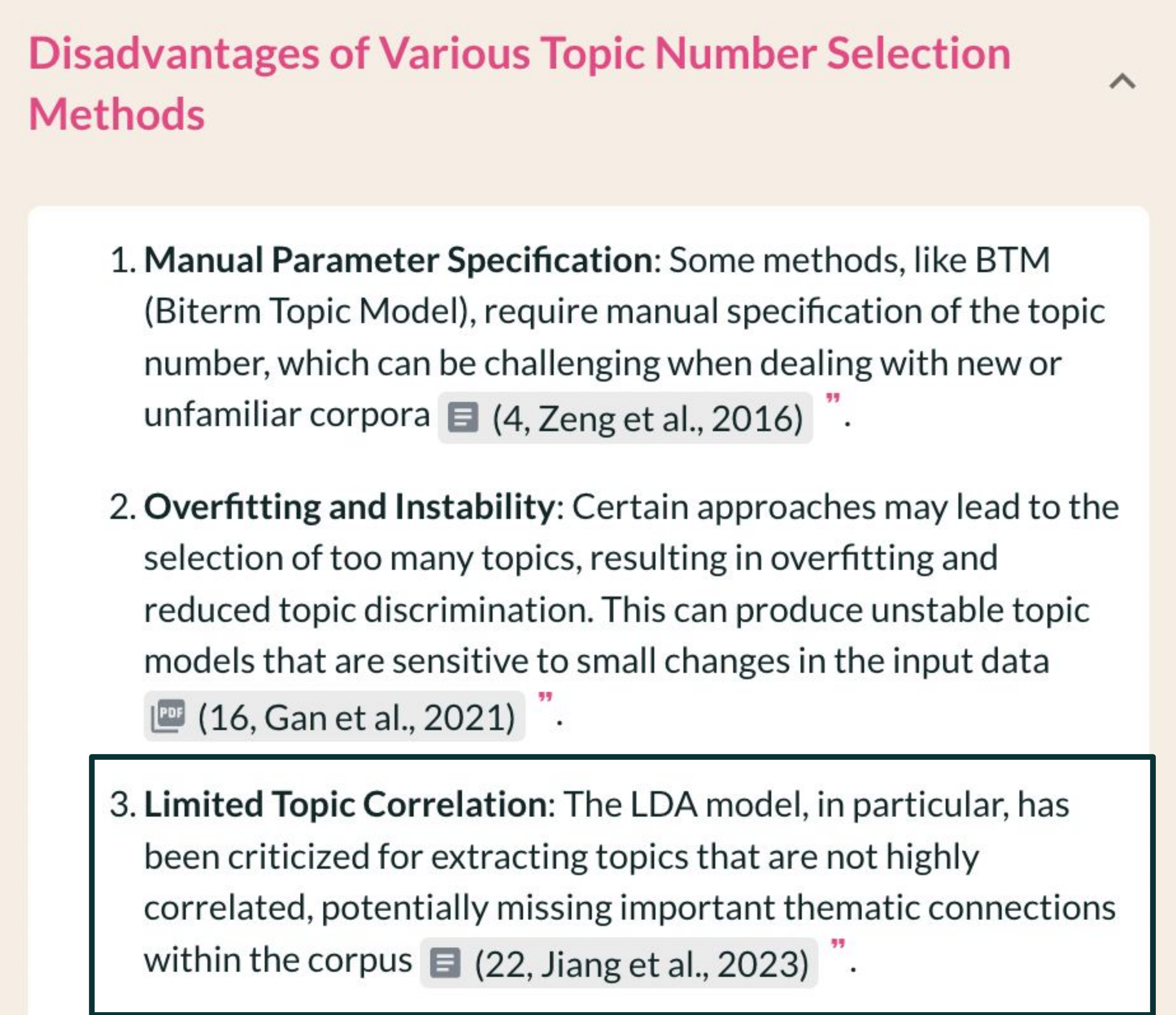


Disadvantages of Various Topic Number Selection Methods

1. **Manual Parameter Specification**: Some methods, like BTM (Biterm Topic Model), require manual specification of the topic number, which can be challenging when dealing with new or unfamiliar corpora (4, Zeng et al., 2016) ”.
2. **Overfitting and Instability**: Certain approaches may lead to the selection of too many topics, resulting in overfitting and reduced topic discrimination. This can produce unstable topic models that are sensitive to small changes in the input data (16, Gan et al., 2021) ”.
3. **Limited Topic Correlation**: The LDA model, in particular, has been criticized for extracting topics that are not highly correlated, potentially missing important thematic connections within the corpus (22, Jiang et al., 2023) ”.

Open Cha

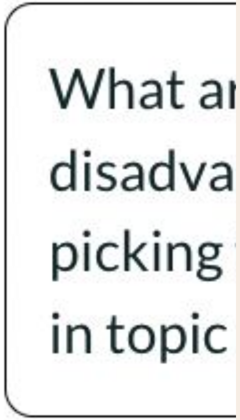
What ar
disadva
picking
in topic



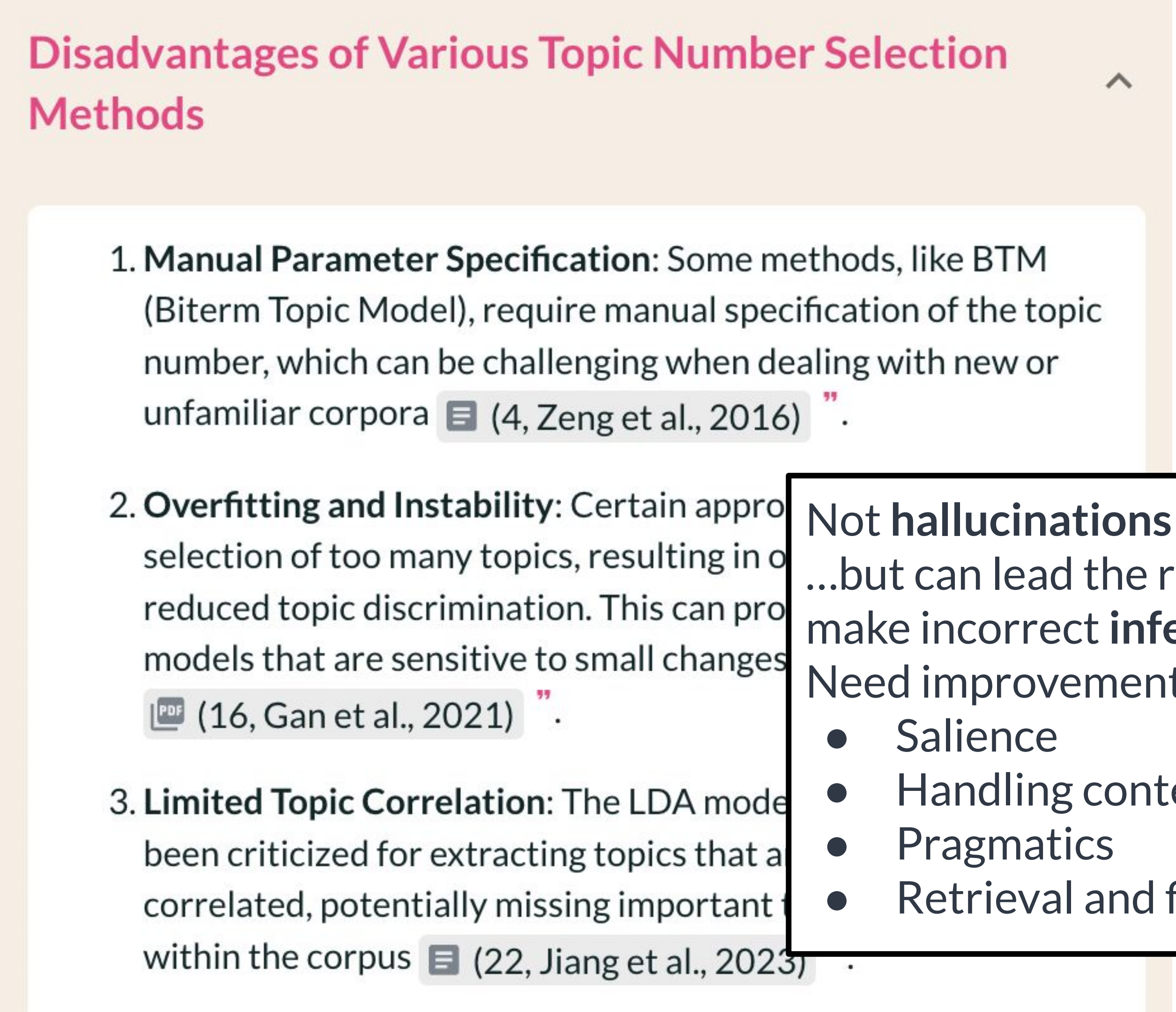
Disadvantages of Various Topic Number Selection Methods

1. **Manual Parameter Specification**: Some methods, like BTM (Biterm Topic Model), require manual specification of the topic number, which can be challenging when dealing with new or unfamiliar corpora (4, Zeng et al., 2016) ”.
2. **Overfitting and Instability**: Certain approaches may lead to the selection of too many topics, resulting in overfitting and reduced topic discrimination. This can produce unstable topic models that are sensitive to small changes in the input data (16, Gan et al., 2021) ”.
3. **Limited Topic Correlation**: The LDA model, in particular, has been criticized for extracting topics that are not highly correlated, potentially missing important thematic connections within the corpus (22, Jiang et al., 2023) ”.

Open Cha



What ar
disadva
picking
in topic

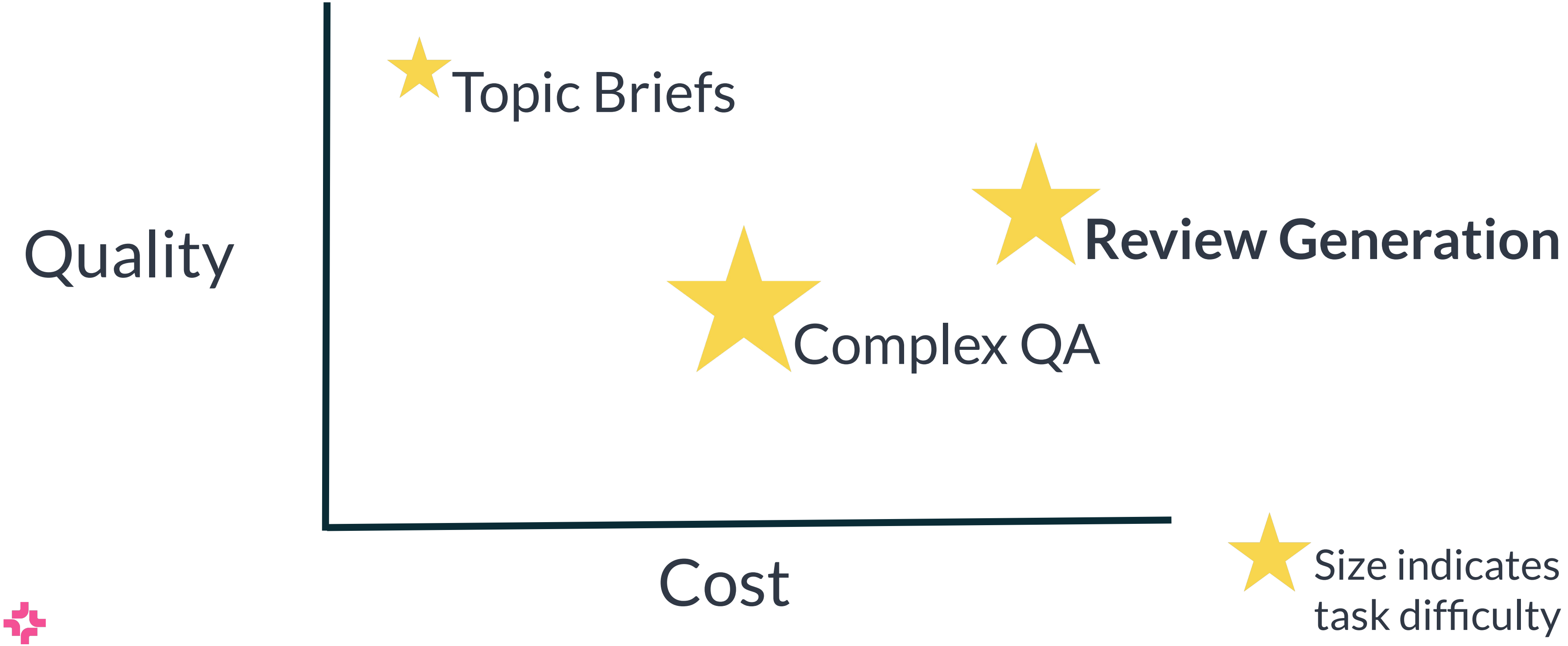


Disadvantages of Various Topic Number Selection Methods

1. **Manual Parameter Specification**: Some methods, like BTM (Biterm Topic Model), require manual specification of the topic number, which can be challenging when dealing with new or unfamiliar corpora (4, Zeng et al., 2016) ”.
2. **Overfitting and Instability**: Certain appro
selection of too many topics, resulting in o
reduced topic discrimination. This can pro
models that are sensitive to small changes
(16, Gan et al., 2021) ”.
3. **Limited Topic Correlation**: The LDA mode
been criticized for extracting topics that a
correlated, potentially missing important
within the corpus (22, Jiang et al., 2023) .

Not **hallucinations**
…but can lead the reader to make incorrect **inferences**
Need improvement in:
- Salience
- Handling context
- Pragmatics
- Retrieval and filtering

Three Tasks, Three Workflows



Review Generation

Larry Birnbaum

Tom Hope

Mike D’Arcy

Given paper, output helpful critiques (as in peer review)

Simple prompts yield unhelpful generic critiques. Solution:

- Taxonomize types of critique (novelty, impact, clarity, experiments, etc.)
    - Ask LLM about each separately

- Handle long input + complexity by decomposing with **multiple agents**, e.g.:
    - One agent reads abstract+intro, outputs **what experiments are necessary** for claims
    - Second agent **checks this** against the paper’s actual experiment section

System Architecture

Retrieval

Dense (GIST embeddings) + Sparse (BM25, SparseEmbed) Vespa index

Fil

Se

su

```
You should start by making a plan of which candidate dimensions
might make sense for the query, ignoring the snippets.
Then, sub-select the dimensions to only retain those that are
actually represented and discussed in the quotes…
```

Clustering

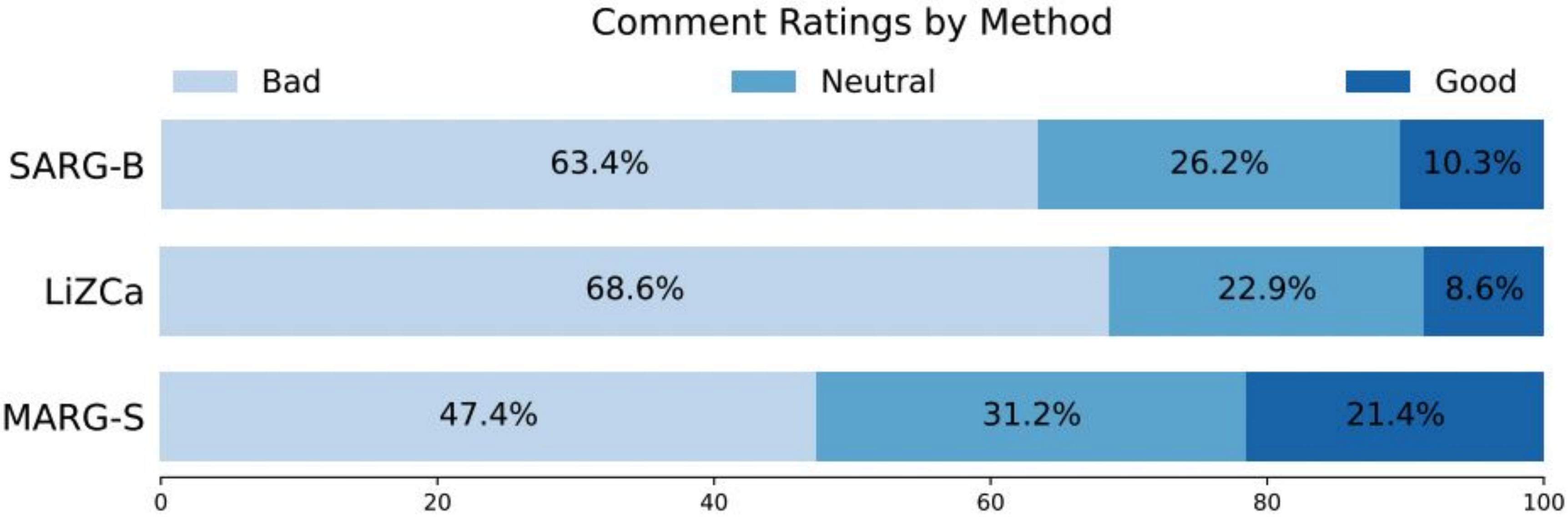
Identify organizing themes, assign snippets to themes

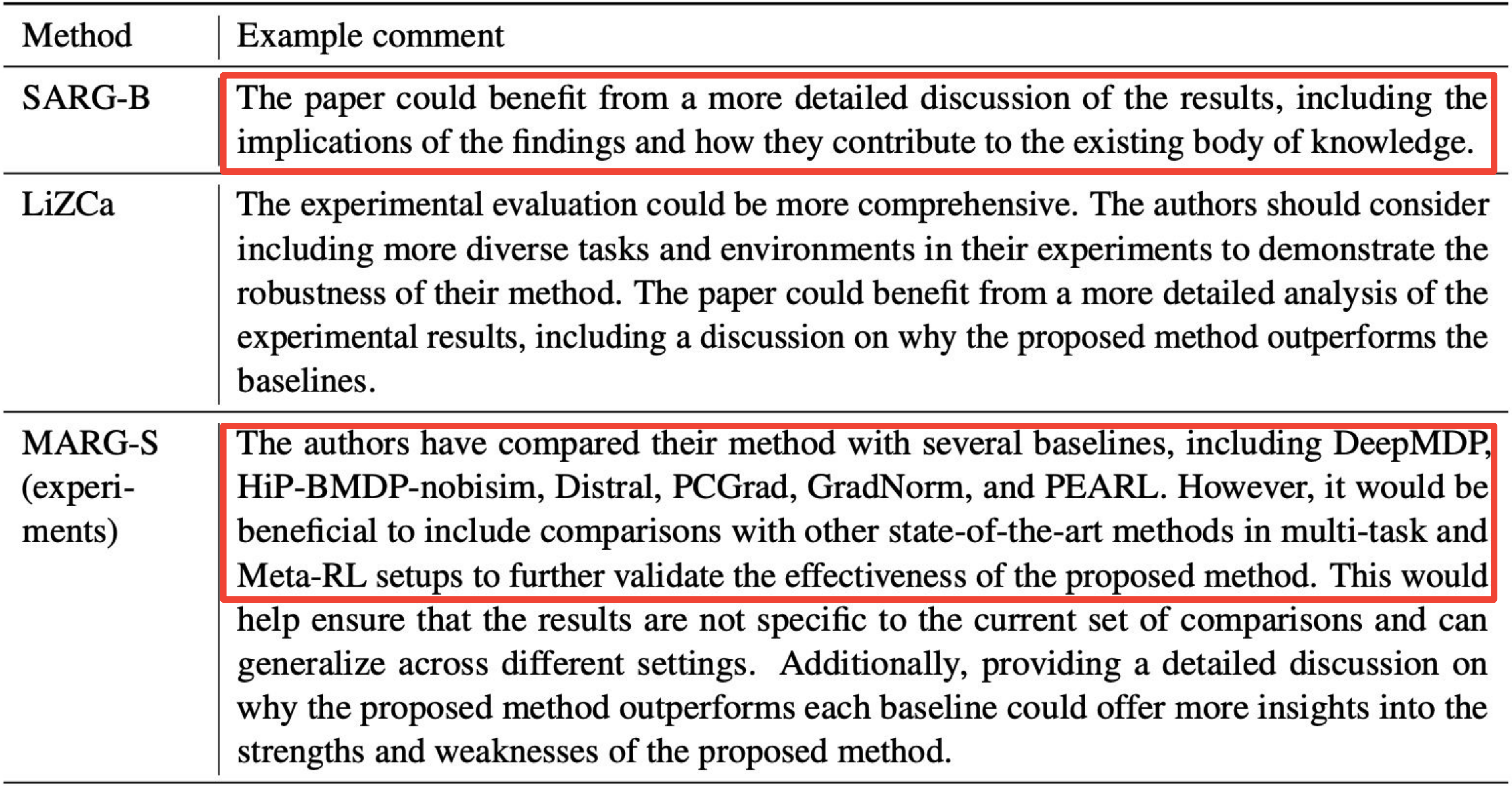
Assembly

Write the response one theme at a time

See also e.g. (Slobodkin et al., 2024)

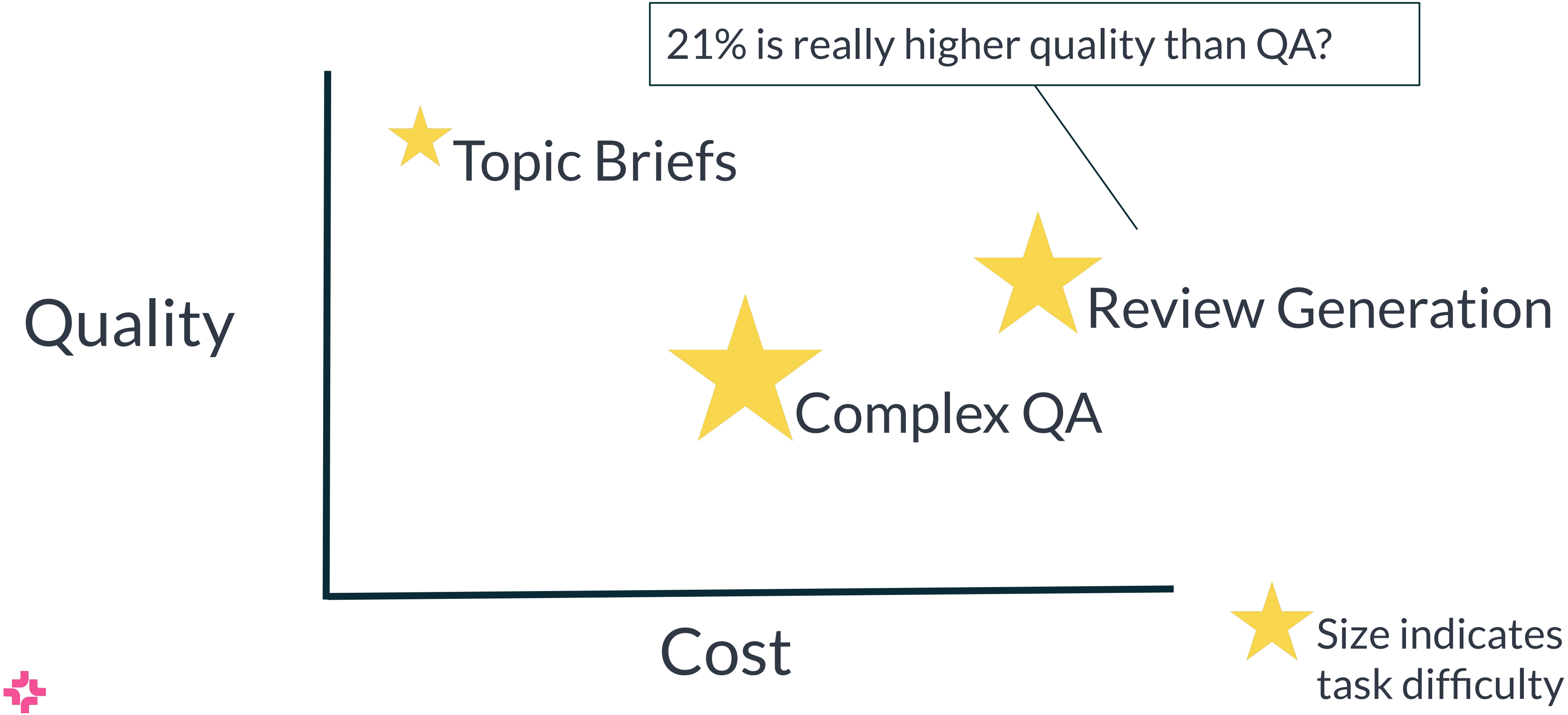
Review Generation Evaluation





Method	Example comment
SARG-B	The paper could benefit from a more detailed discussion of the results, including the implications of the findings and how they contribute to the existing body of knowledge.
LiZCa	The experimental evaluation could be more comprehensive. The authors should consider including more diverse tasks and environments in their experiments to demonstrate the robustness of their method. The paper could benefit from a more detailed analysis of the experimental results, including a discussion on why the proposed method outperforms the baselines.
MARG-S (experiments)	The authors have compared their method with several baselines, including DeepMDP, HiP-BMDP-nobisim, Distral, PCGrad, GradNorm, and PEARL. However, it would be beneficial to include comparisons with other state-of-the-art methods in multi-task and Meta-RL setups to further validate the effectiveness of the proposed method. This would help ensure that the results are not specific to the current set of comparisons and can generalize across different settings. Additionally, providing a detailed discussion on why the proposed method outperforms each baseline could offer more insights into the strengths and weaknesses of the proposed method.

Three Tasks, Three Workflows



21% is really higher quality than QA?

Topic Briefs

Review Generation

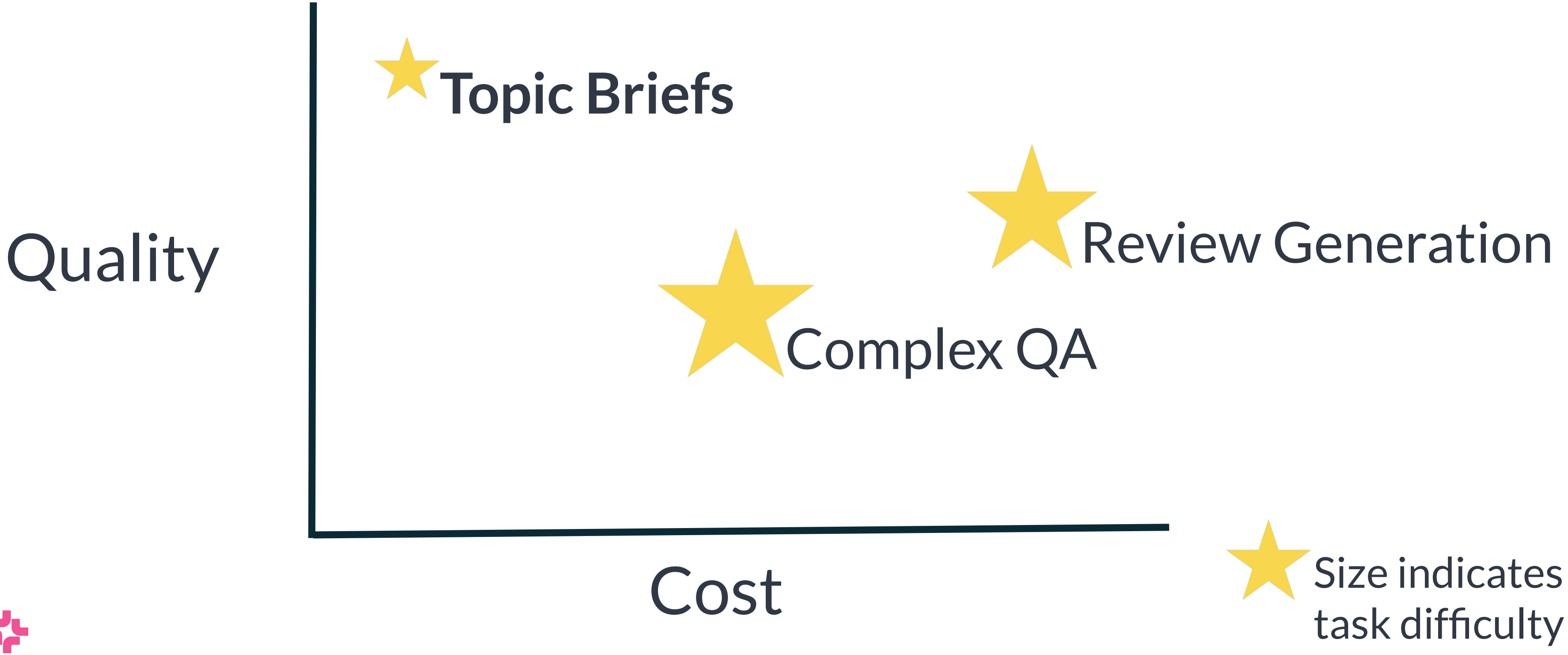
Complex QA

Quality

Cost

Size indicates task difficulty

Three Tasks, Three Workflows



Scientific Topic Pages

- Most requested Semantic Scholar feature in user surveys
  - A Wikipedia for the “long tail” of science
    - E.g., RoBERTa, BERTScore, Transformer-XL, SpanBERT—all have 1000+ citations, yet no Wikipedia page
  - Allow users to quickly get short descriptions of topics, and discover papers on those topics

Amanpreet Singh

Luca Soldaini

Pao Siangliulue

Sergey Feldman

Scientific Topic Pages

- Most requested Semantic Scholar feature in user surveys
  - A Wikipedia for the “long tail” of science
    - E.g., RoBERTa, BERTScore, Transformer-XL, SpanBERT—all have 1000+ citations, yet no Wikipedia page
  - Allow users to quickly get short descriptions of topics, and

Focus on a **simplified** task: short (two-sentence) topic descriptions and links to foundational + recent papers

Amanpreet Singh

Luca Soldaini

Pao Siangliulue

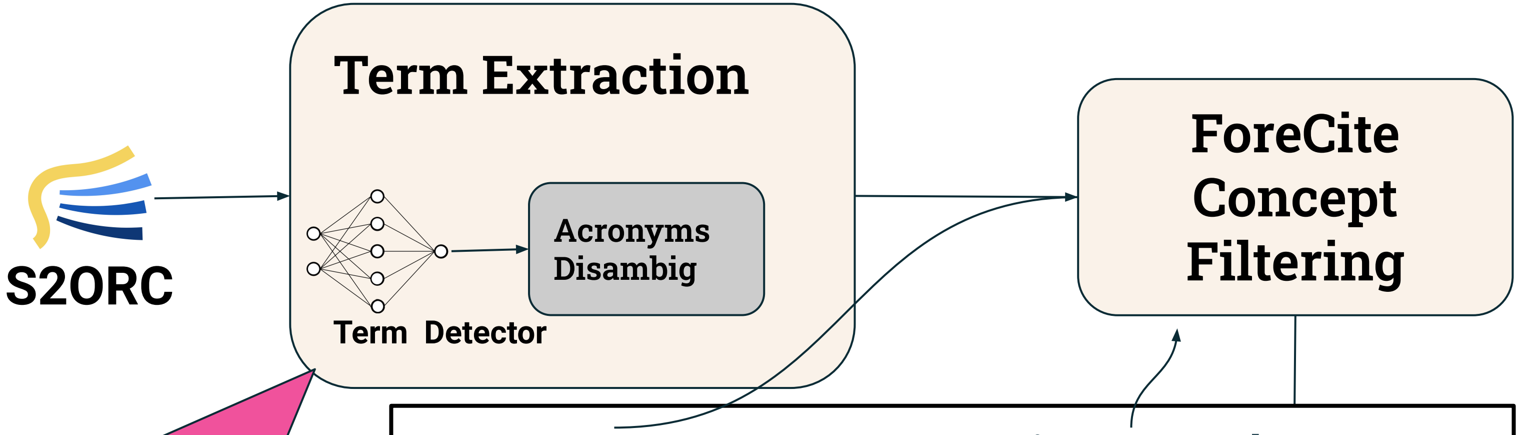
Sergey Feldman

Demo

Key Challenges

- What are the concepts?
- How to generate accurate descriptions?

What are the concepts?



~~SciSpacy~~
RoBERTa trained on 100k silver examples from GPT

We use **CNNs** to process our **input graph**.

ForeCite answers: What's a "concept"?

- **Input:** corpus of scientific papers and their candidate terms
- **Output:** the subset of candidate terms that are **scientific concepts**

LLaMA ✓
gradient penalty ✓
asynchronous advantage actor-critic ✓

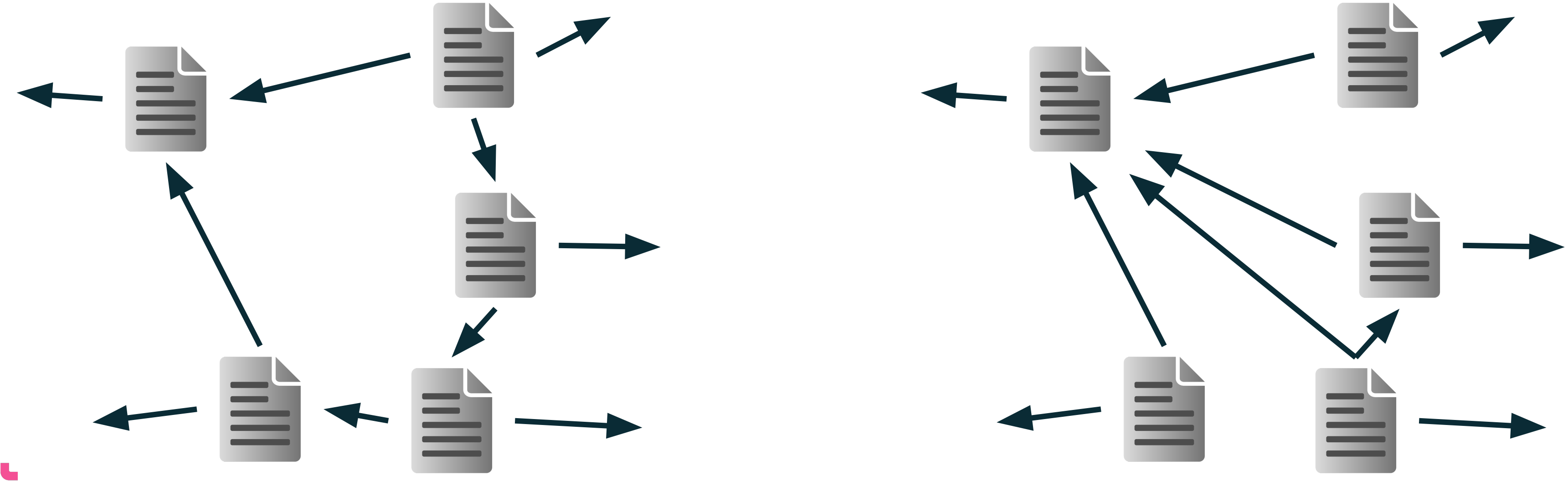
popular model ✗
input graph ✗

Daniel King

“We use BERT (Devlin et al. 2018) as our base model…”

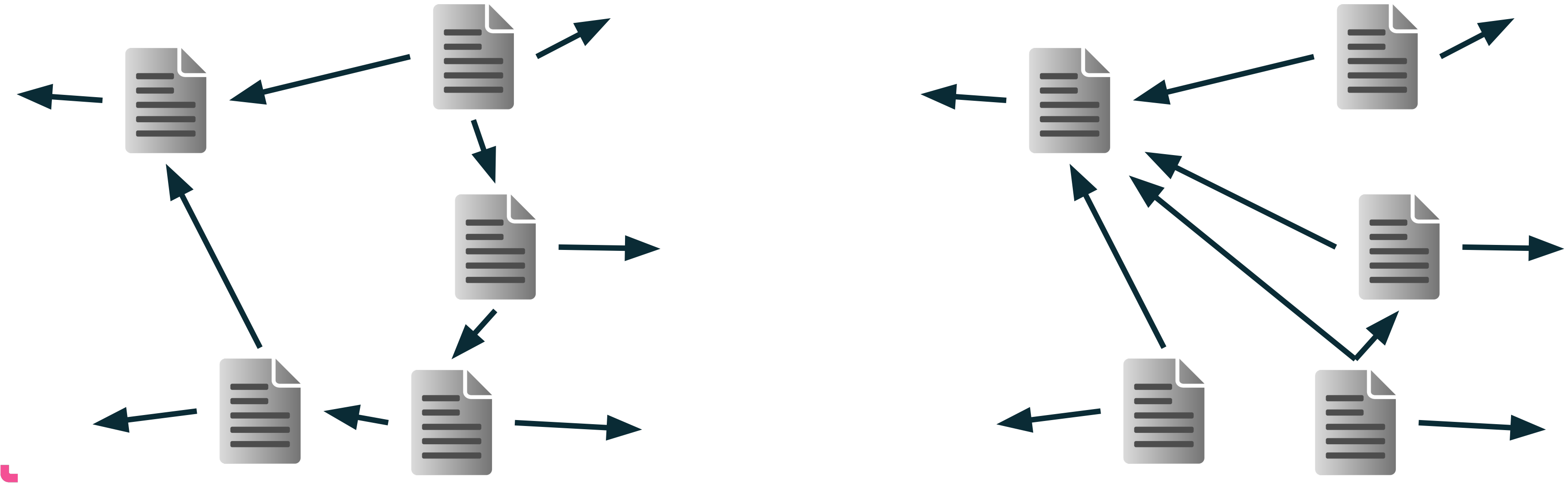
Finding Concepts from Phrase Subgraphs

- Nodes = Papers Edges = Citations Subgraph for a term ***t*** = all papers that contain ***t***



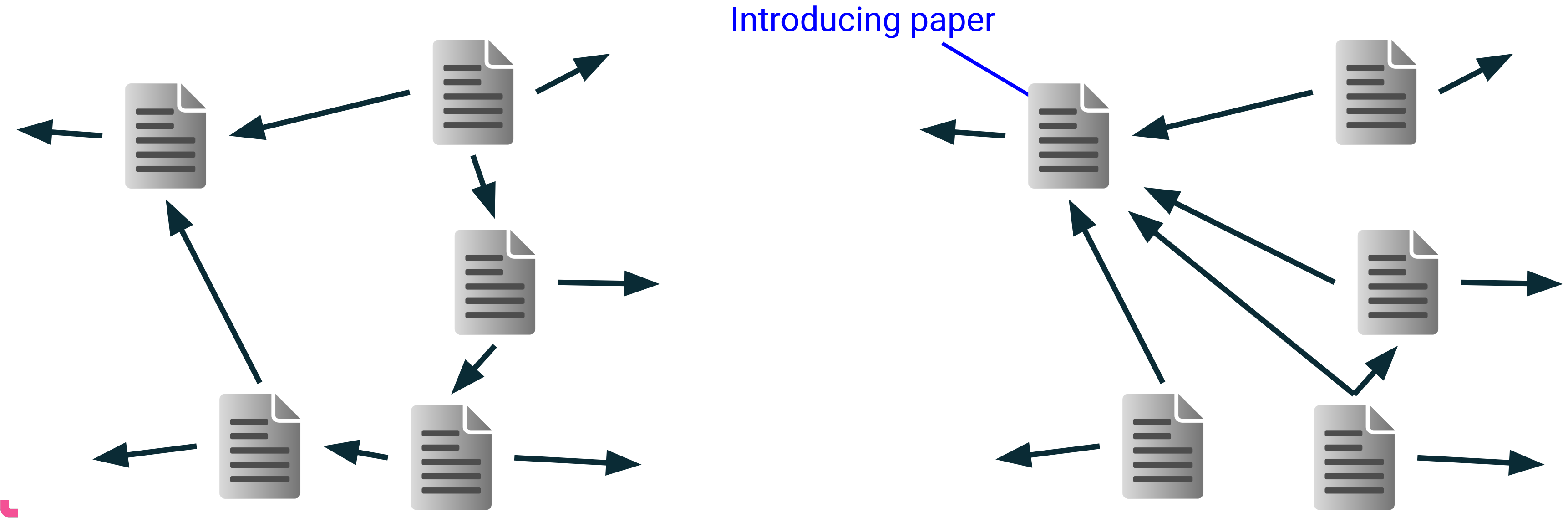
Finding Concepts from Phrase Subgraphs

- Nodes = Papers   Edges = Citations   Subgraph for a term ***t*** = all papers that contain ***t***
- Prior work: ***t*** is more likely a concept when the subgraph for ***t*** is ***more dense***



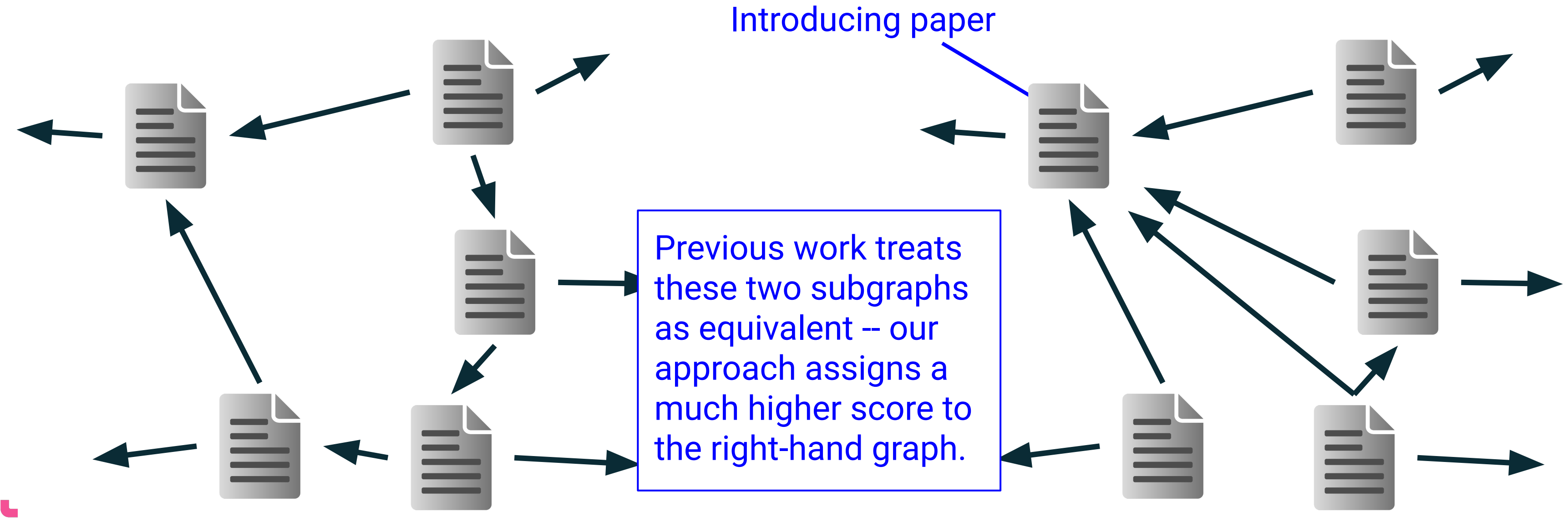
Finding Concepts from Phrase Subgraphs

- Nodes = Papers Edges = Citations Subgraph for a term ***t*** = all papers that contain ***t***
- Prior work: ***t*** is more likely a concept when the subgraph for ***t*** is ***more dense***
- Our work: ***t*** is more likely a concept when the subgraph for ***t*** has an ***introducing paper***



Finding Concepts from Phrase Subgraphs

- Nodes = Papers Edges = Citations Subgraph for a term ***t*** = all papers that contain ***t***
- Prior work: ***t*** is more likely a concept when the subgraph for ***t*** is ***more dense***
- Our work: ***t*** is more likely a concept when the subgraph for ***t*** has an ***introducing paper***



Introducing paper

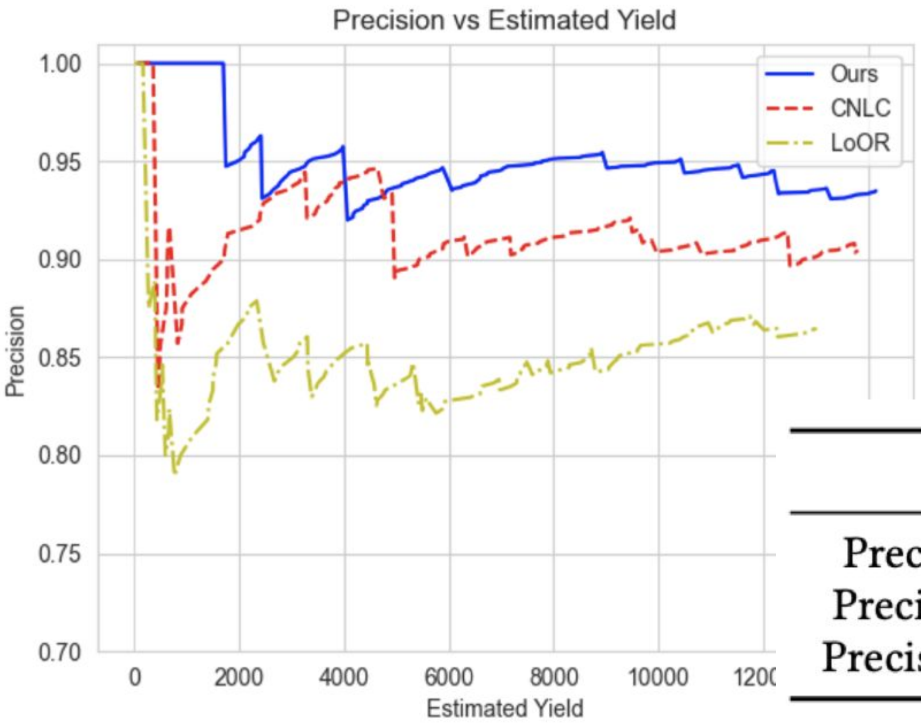
Previous work treats these two subgraphs as equivalent -- our approach assigns a much higher score to the right-hand graph.

ForeCite

ForeCite(term t) =

$\max_p$ P(cite paper p | contain t) * lg (1 + #papers containing t and citing p)

Results



	FORECITE	CNLC	LoOR
Precision@100	**1**	0.93	0.91
Precision@1000	**0.99**	0.86	0.81
Precision@10000	0.93	0.88	0.84

Results - top 5

LoOR	CNLC	**ForeCite**
codeword	VQA	**fast gradient sign method**
received signal ✗	adversarial example	**DeepWalk**
achievable rate	adversarial perturbation	**BERT**
convolutional layer	ImageNet	**node2vec**
antenna	person re-identification	**region proposal network**

Generating descriptions

For less common knowledge, require stronger evidence from literature:

- For **rare concepts** (occurs in < 1000 titles/abstracts) require LLM output that **draws directly from the paper with highest ForeCite score**
- For more **common concepts**, just provide top-ForeCite-score papers as **context** to the LLM

Lots of manual prompt engineering

Generating descriptions

For less common knowledge, require stronger evidence from literature:

```
You are Neil DeGrasse Tyson, an expert scientific communicator.
You have been tasked with summarizing information about topics, which
you do very well.
```

- that
- **context** to the LLM

Lots of manual prompt engineering

Generating descriptions

For less common knowledge, require stronger evidence from literature:

- that

```
You are Neil DeGrasse Tyson, an expert scientific communicator.
You have been tasked with summarizing information about topics, which
you do very well.
```

- **context** to the LLM

Lo

```
...
- Exclude facts that refer to external sources such as figures,
references or other sources not included
- Exclude math facts
- Exclude citations
- Exclude numerical results
```

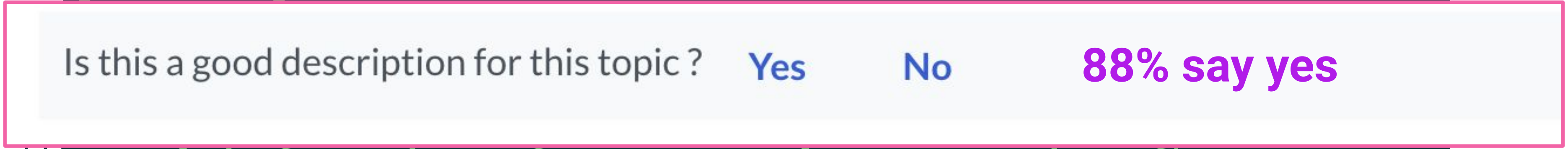
Generating descriptions

For less common knowledge, require stronger evidence from literature:

- that

L

```
You are Neil DeGrasse Tyson, an expert scientific communicator.
You have been tasked with summarizing information about topics, which
you do very well.
```

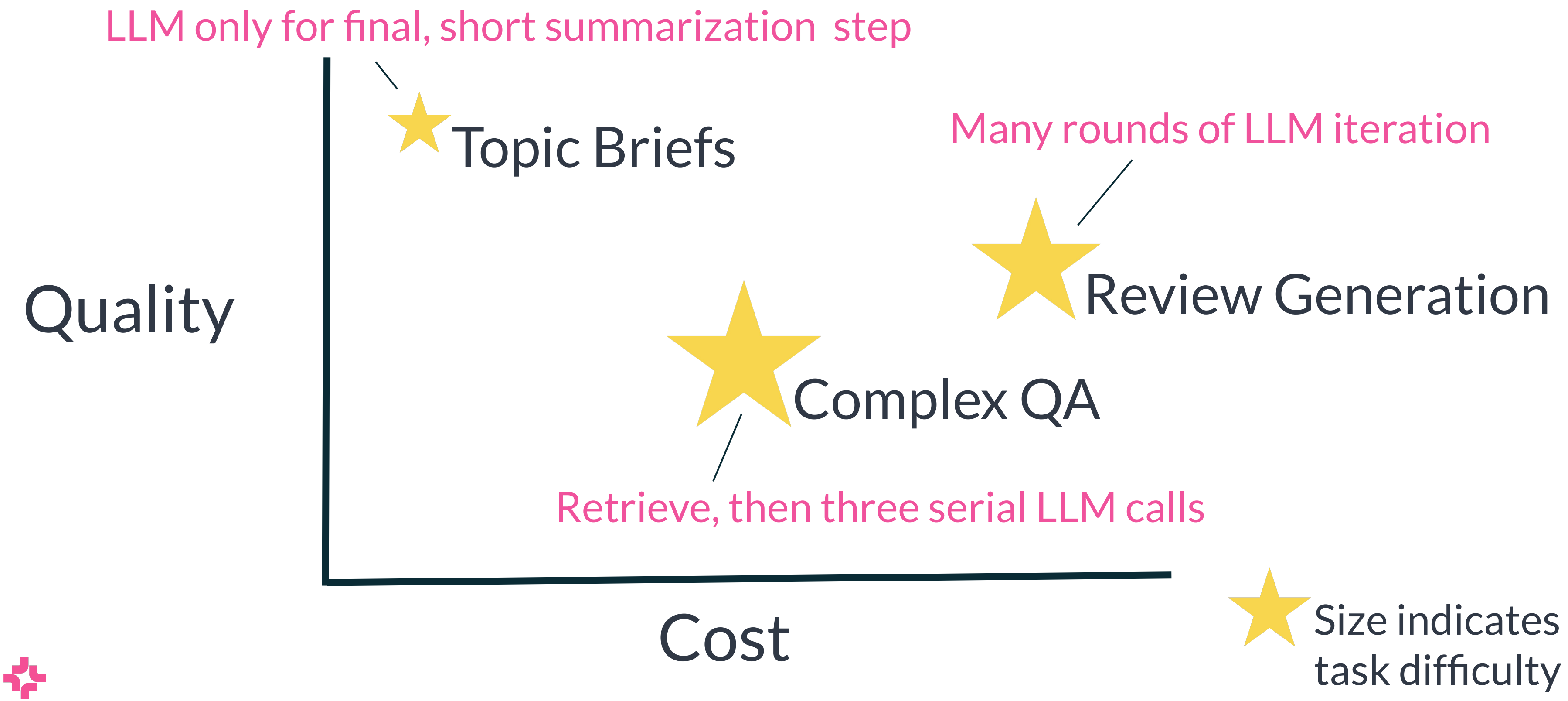


Is this a good description for this topic ?

88% say yes

```
- Exclude facts that refer to external sources such as figures,
references or other sources not included
- Exclude math facts
- Exclude citations
- Exclude numerical results
```

Conclusions – three different LLM workflows


LLM only for final, short summarization step
Topic Briefs
Many rounds of LLM iteration
Quality
Review Generation
Complex QA
Retrieve, then three serial LLM calls
Cost
Size indicates task difficulty

Conclusions

Lessons:

- LLMs are powerful, but don't be afraid to radically **simplify** the task
- Look for applications with **low cost of error**
- Having LLM **delay consideration** of retrieved content is helpful
    - cf. (Grunde-McLaughlin et al., 2023)

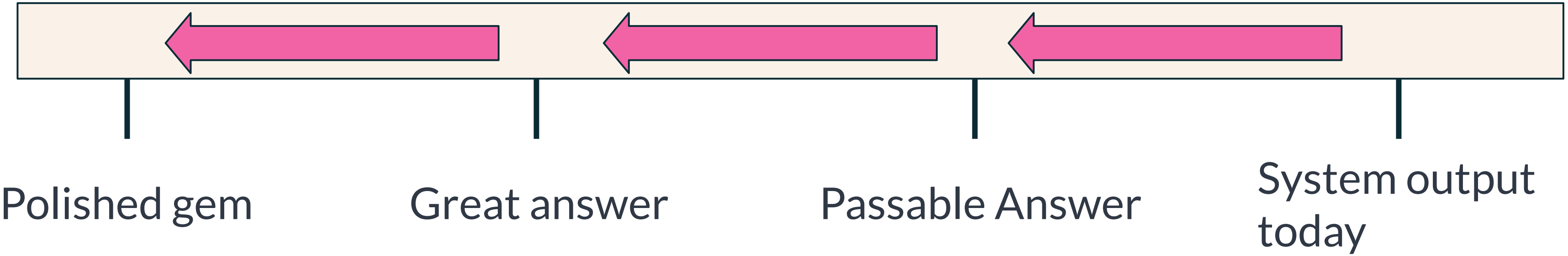
Many Challenges Remain, e.g. evaluation

(Small sampling of) Opportunities

- Understanding the User’s Context
- Proactivity

(Small sampling of) Opportunities

- Understanding the User's Context
- Proactivity



Thank you

Luca Soldaini

Tal August

Kyle Lo

Amanpreet Singh

Joseph Chee Chang

Dave Wadden

Pao Siangliulue

Sergey Feldman

Dan Weld

Arman Cohan

Lucy Lu Wang

Amy X. Zhang

Aakanksha Naik

Tom Hope

Maria Antoniak

Jonathan Bragg