Bridging the Evaluation Gap:
Leveraging Large Language Models for Topic Model Evaluation

Zhiyin Tan *L3S Research Center, Leibniz University Hannover, Hannover, Germany*

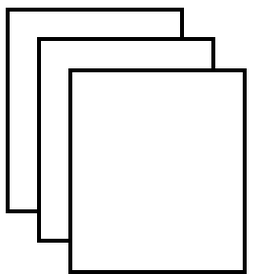
Jennifer D'souza *TIB Leibniz Information Centre for Science and Technology, Hannover, Germany*

21st Conference on Information and Research Science
Connecting to Digital and Library Science

February 20-21 2025
Udine, Italy

1. A Brief Introduction of Topic Modeling

A unsupervised machine learning method used to discover hidden theme within a collection of documents.



Collection of Documents

Topic Modeling

Topic 1: words
Topic 2: words
Topic 3: words
…

Document 1: topics
Document 2: topics
Document 3: topics
…

Topic-Word Distribution

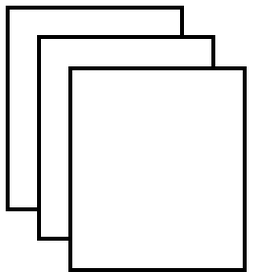
the probability distribution of words for each topic, indicating the most relevant words that define a topic.

Document-Topic Distribution

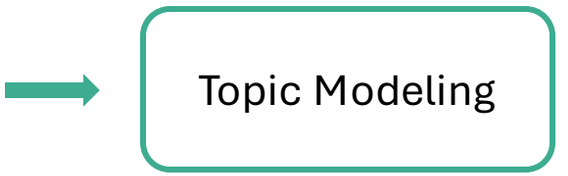
the probability distribution of topics for each document, showing how much each topic contributes to a given document.

1. A Brief Introduction of Topic Modeling

Topic-Word Distribution: each topic consists of a list of topic words.



Collection of Documents



Topic	Topic Words Distribution ϕ			
	w_1	w_2	...	w_V
t_1	$\phi_{1,1}$	$\phi_{1,2}$	...	$\phi_{1,V}$
...	...	...	...	...
t_K	$\phi_{K,1}$	$\phi_{K,2}$		$\phi_{K,V}$

Examples
(**K**=3, **N**=3)

Topic 1: desert, cake, apple

Topic 2: pen, pens, notebook

Topic 3: notebook, apple, stock

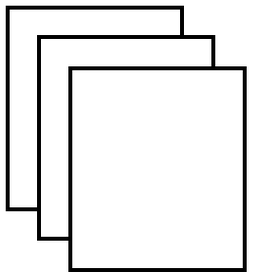
Topic: $T = \{t_1, t_2, \ldots, t_K\}$, $|T|=K$, where **K is the number of topics**.

Topic words: $W = \{w_1, w_2, \ldots, w_V\}$, $|W|=V$, where V is the vocabulary size.

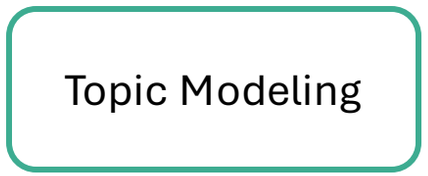
Number of **top topic words** retained for each topic: **N**

1. A Brief Introduction of Topic Modeling

Document-Topic Distribution: each document will be assigned a probability for each topic.



Collection of Documents



Document	Document-Topic Distribution θ			
	t_1	t_2	...	t_K
d_1	$\theta_{1,1}$	$\theta_{1,2}$	...	$\theta_{1,K}$
...	...	...	...	...
d_M	$\theta_{M,1}$	$\theta_{M,2}$	...	$\theta_{M,K}$

Examples
(K=2, Top 1 Topic)

Document 1: Topic 2

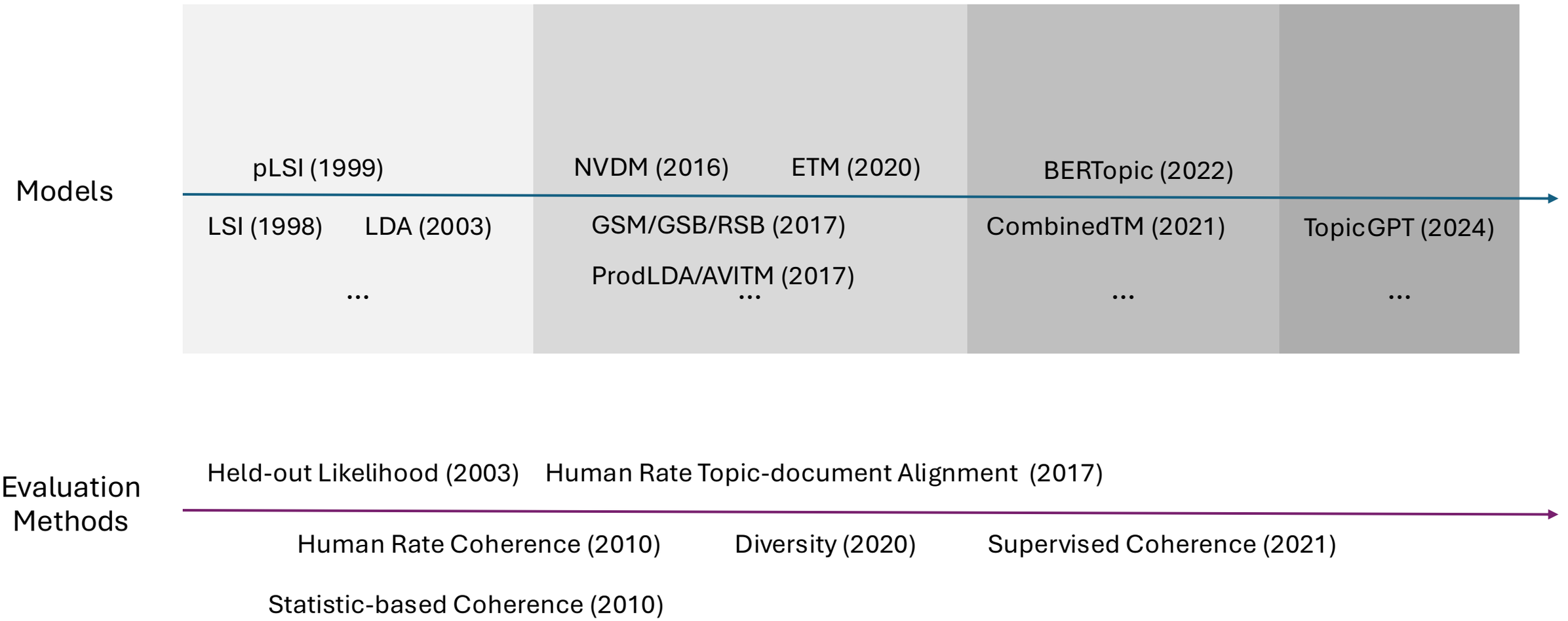
Document 2: Topic 1

Document 3: Topic 2

Topic: $T = \{t_1, t_2, \ldots, t_K\}$, $|T|=K$, where K is the number of topics.

Documents: $D = \{d_1, d_2, \ldots, d_M\}$, $|D|=M$, where M is the number of documents.

1. A Brief Introduction of Topic Modeling



1. A Brief Introduction of Topic Modeling

"...should thus focus on evaluations that depend on real-world task performance rather than optimizing likelihood-based measures."

-- *Reading tea leaves: How humans interpret topic models, 2009*

"There is a disconnect between how topic models are evaluated and why we expect topic models to be useful."

-- *Probabilistic topic models, 2012*

1. A Brief Introduction of Topic Modeling

Applications

Extracting Themes & Trends

Identify key themes in documents (e.g. news, research papers, reviews, and social media posts) and track topic evolution over time.

Topic Words Quality

Document Categorization

Group documents into topic-based categories for efficient organization.

Distinctive Topic

Content Recommendation

Suggest relevant content by aligning documents with similar topics.

Topic-document Alignment Quality

2. Challenges of Topic Modeling Evaluation

Topic Words Quality

Example

Topic 1: desert, cake, apple

Statistic-based Coherence Metrics (e.g., UMass, UCI, NPMI, CV):
Measure how well topic **words co-occur** in actual text **corpora**.

Depending on corpus selection and may not generalize well across domains.

Rated Coherence:
Rate topic words on an ordinal scale **by human or LLMs**.

Lack of interpretability of the rate.

Supervised Comparison:
Compare topic words with human-defined or LLM-generated **topic labels**.

Rely on the quality of predefined labels.

Solutions: enhance **evaluation interpretability** by integrating **outlier detection** and **duplicate concept detection** to identify incoherent or redundant topic words.

2. Challenges of Topic Modeling Evaluation

Distinctive Topic

Example

Topic 1: desert, cake, ***apple***

Topic 2: pen, pens, ***notebook***

Topic 3: ***notebook, apple****, stock*

Statistical Diversity Metrics (e.g., uniqueness, redundancy, overlap):
Word-level comparison across topics.

Overly strict penalization of words that have different meanings in different contexts (co-occurring topic words).

Embedding-Based Diversity:
Measure topic similarity using **embeddings**.

Rely on pre-trained knowledge.

Solution: enhance **context-aware, nuanced semantic pairwise topic comparison** by leveraging LLMs.

2. Challenges of Topic Modeling Evaluation

Topic-document

Alignment Quality

Human Rate:
Rate the relevance of each topic to a given document on an ordinal scale by human.

Lack of interpretability of the rate.
Human resource consuming, especially for long documents

Example

Topic 3: ***notebook, apple****, stock*

Document 8:

OLEDs are already in use in ***notebooks*** *in the Windows world for years, but* ***Apple*** *isn't satisfied with the quality.*

Solutions: enhance **evaluation interpretability** by **granularly comparing topic-document pairs**, i.e., detecting whether a topic **over-represents** or **under-represents** the document it is aligned with.

3. An Unified LLM-based Evaluation Framework

Topic Words Quality

Coherence:

1. Rating [1-3] ↑ C_{rate}
2. Outlier detection $C_{outlier}$

A topic word is outlier, if it has been identified ≥ 3 times out of 5 times. Record 1) all outliers 2) total number of outliers for one topic.

Adversarial Test

Manually add an outlier to each topic word list and count the number of times the LLM detects that outlier.

Repetitiveness:

1. Rating [1-3] ↑ R_{rate}
2. Duplicate concept detection $R_{duplicate}$

A topic word is labelled 1 if it has ≥ 1 conceptual duplicates in the topic word list, and 0 otherwise. Record 1) duplicate pairs 2) total number of the duplicate pairs for one topic.

Adversarial Test

Randomly select a word w from the topic word list and manually pair it with a conceptually similar word s. Calculate the number of times the LLM detects its conceptually similar word s for a given w.

Distinctive Topic

Diversity:

1. Rating [1-3] ↑ D_{rate}

Topic-document

Alignment Quality

Topic-document Alignment:

1. Irrelevant Topic Words Detection $A_{\text{ir-topic}}$

Record 1) the topic words that are irrelevant to the aligned document 2) number of irrelevant topic words for a topic-document pair

2. Missing Themes Detection $A_{\text{missing-theme}}$

Record 1) the themes present in the document that are not included in the topic word list 2) number of these themes for a topic-document pair.

4. Experiments

Datasets

	20 Newsgroups	AGRIS
Documents	16,309	50,067
Tokens	783,151	14,370,425
Vocabulary	1583	131,521 (top 10,000)

LLMs

Mistral	Mistral-7B-Instruct-v0.3
Llama	Meta-Llama-3.1-8B-Instruct
Qwen	Qwen2.5-14B-Instruct

Topic models

	Hyperparameters
LDA	11
ProdLDA	4
CombinedTM	4
BERTopic	4

- Number of topic K: 50, 100
- Top N topic words: 10

Evaluation Metrics

- LLM-based metrics
- Statistic-based: C_v (Coherence), D_{unique}(Diversity)

(Running multiple times, selecting hyperparameter combinations for each topic model based on model performance evaluated by two statistics-based metrics)

5. Results

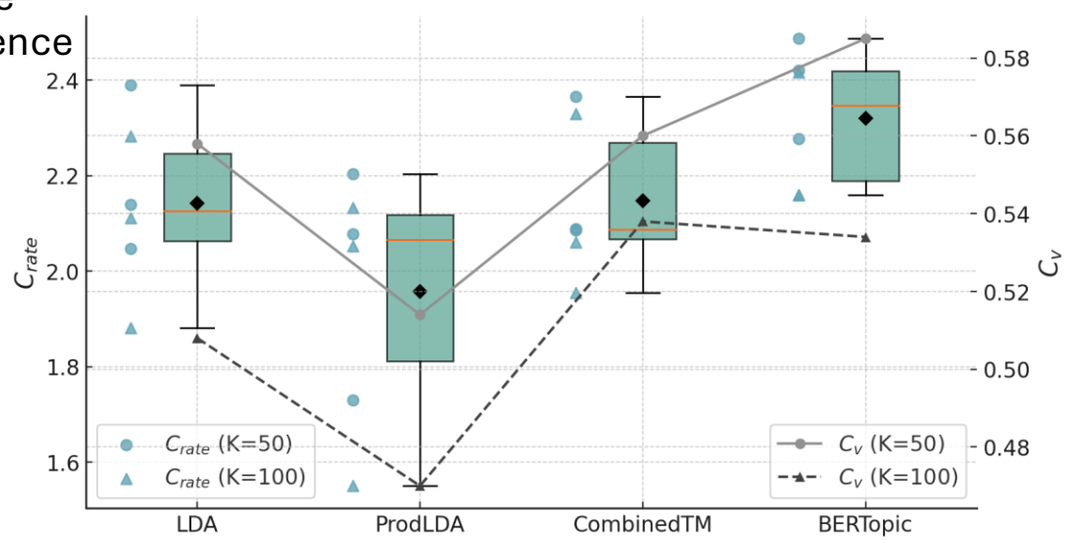
Adversarial Test

20NG	Mistral	Llama	Qwen
$AdvT_{\text{outlier}}$	0.77	0.81	0.90
$AdvT_{\text{duplicate}}$	**0.37**	0.81	0.84

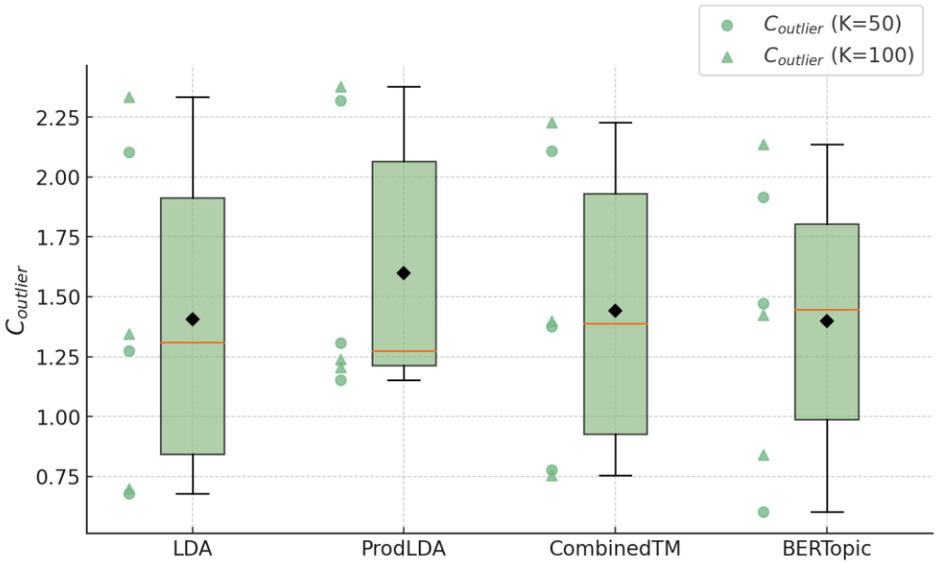
AGRIS	Mistral	Llama	Qwen
$AdvT_{\text{outlier}}$	0.82	0.85	0.93
$AdvT_{\text{duplicate}}$	**0.29**	0.74	0.81

5. Results – Coherence & Repetitiveness – 20NG

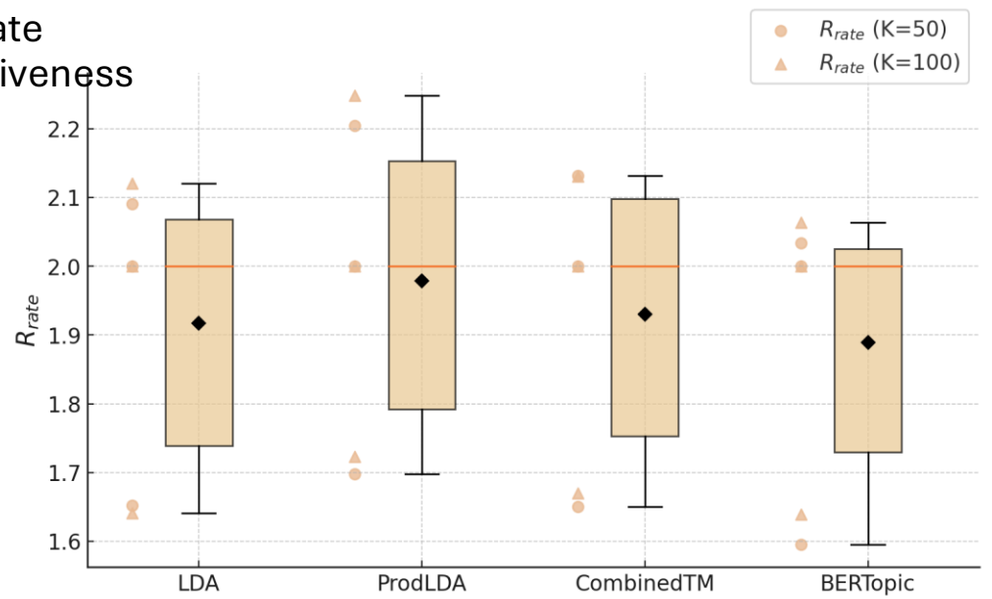
Rate Coherence



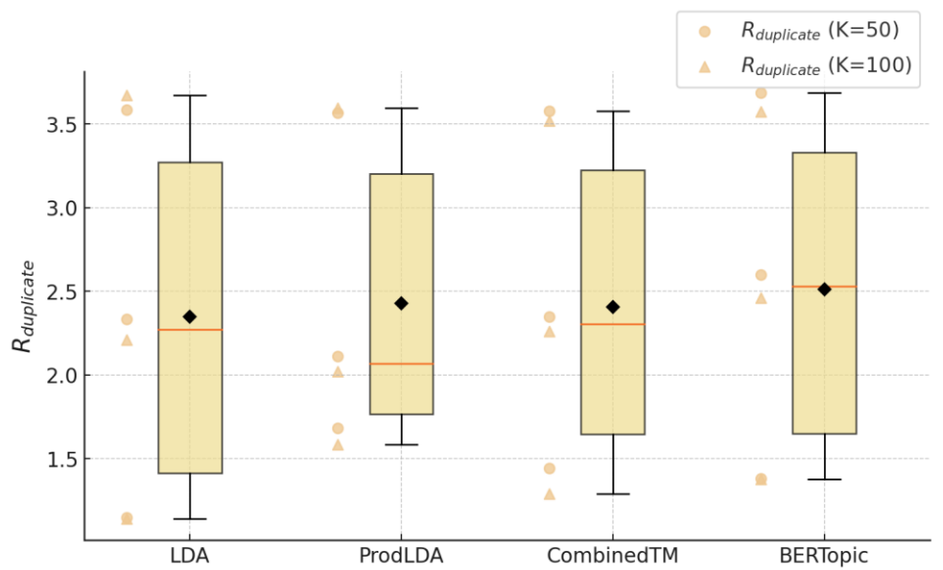
Outlier Detection



Rate Repetitiveness

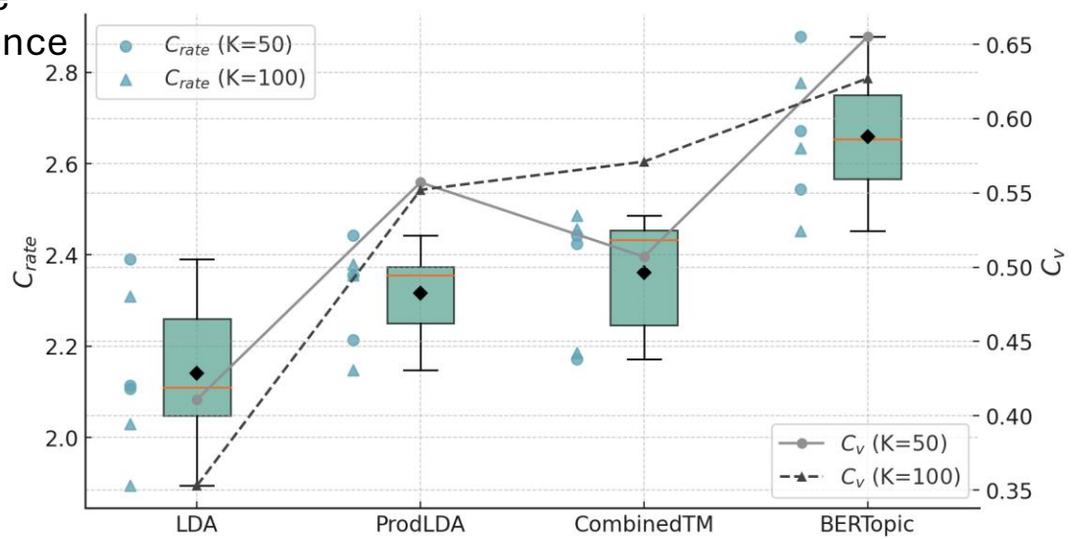


Duplicate Concept Detection

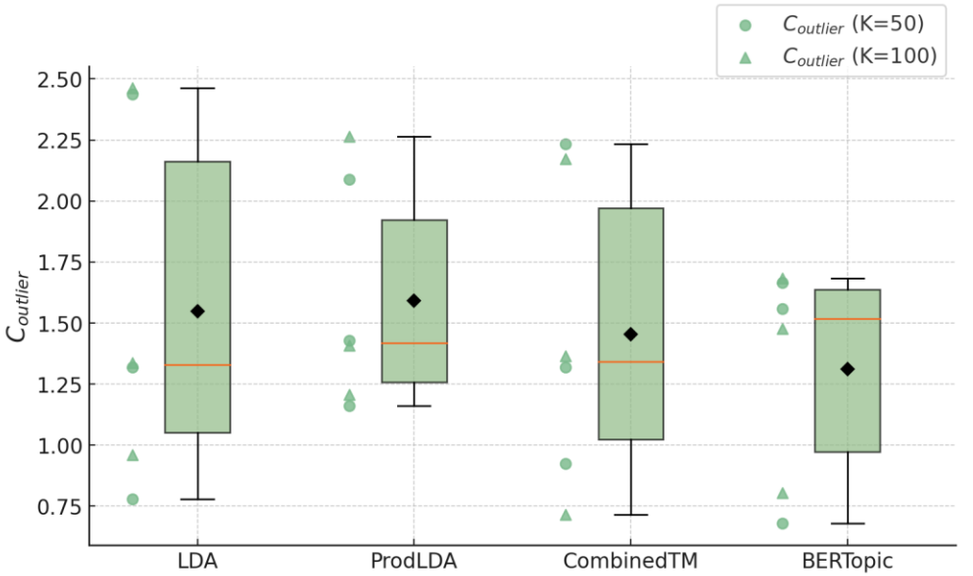


5. Results – Coherence & Repetitiveness - AGRIS

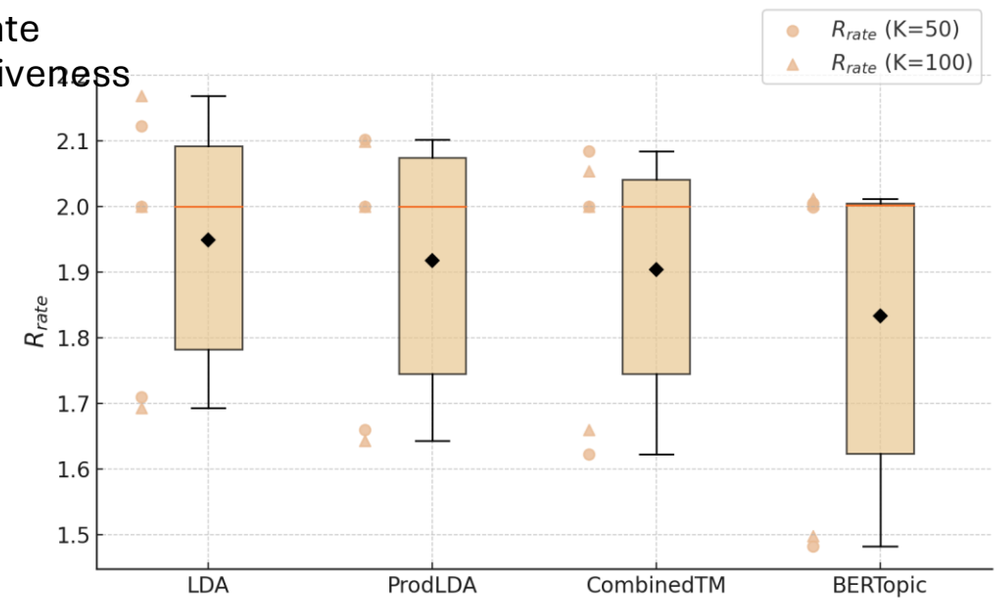
Rate Coherence



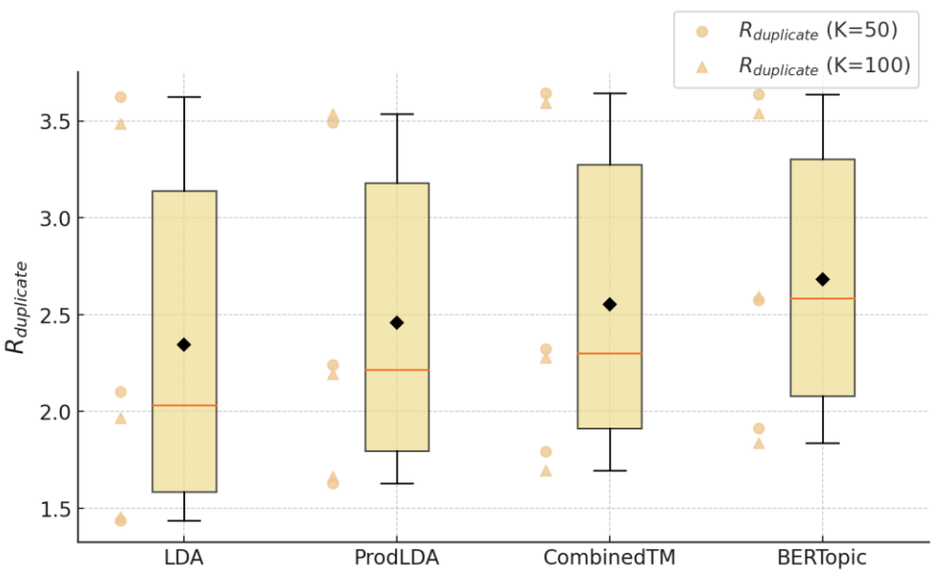
Outlier Detection



Rate Repetitiveness

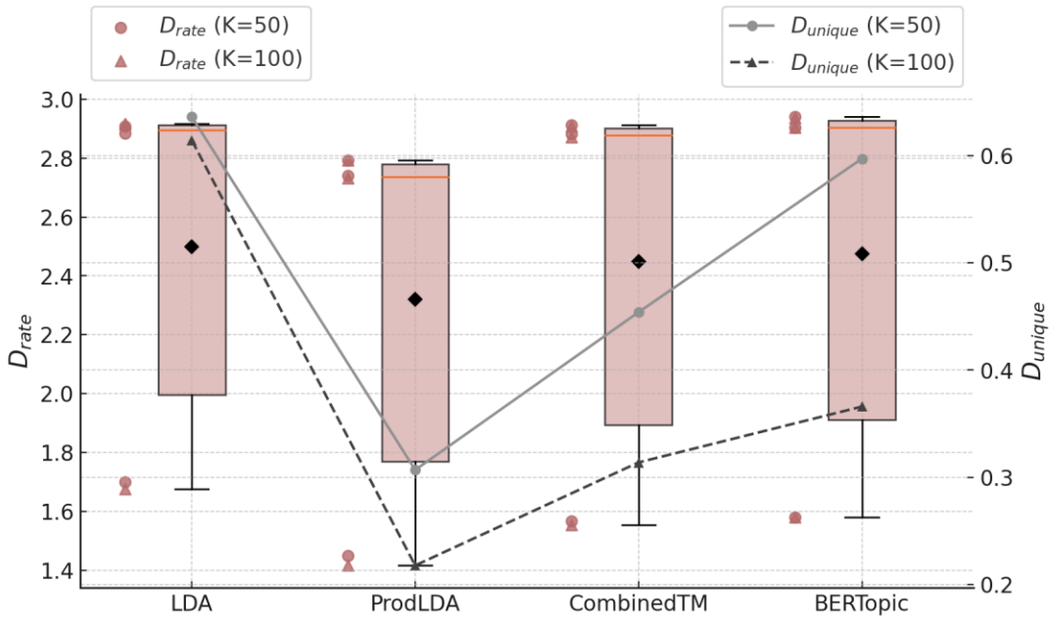


Duplicate Concept Detection

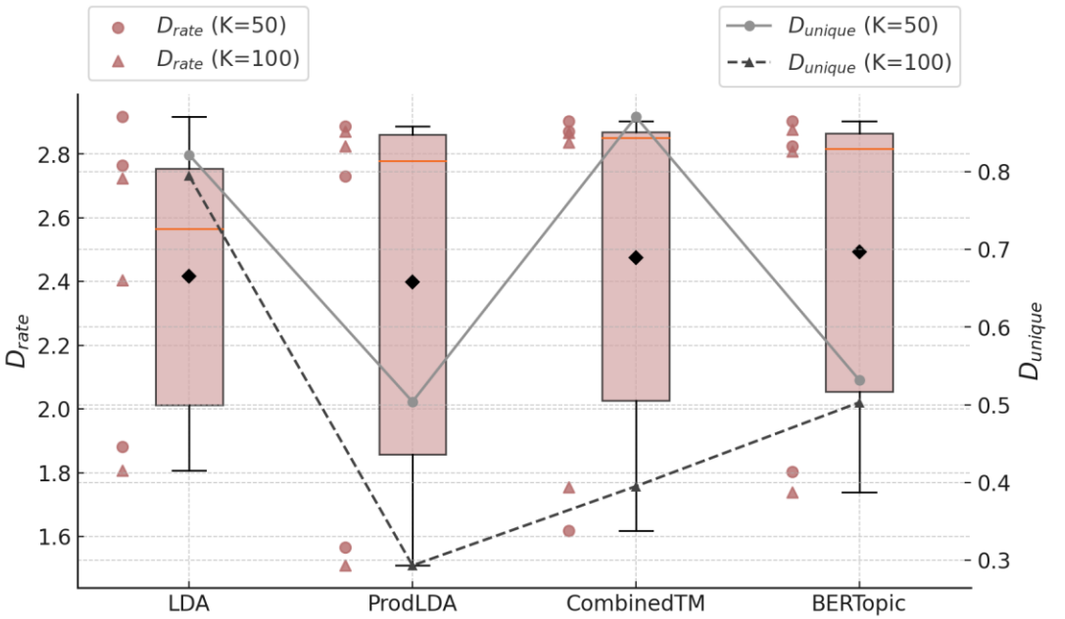


5. Results – Diversity

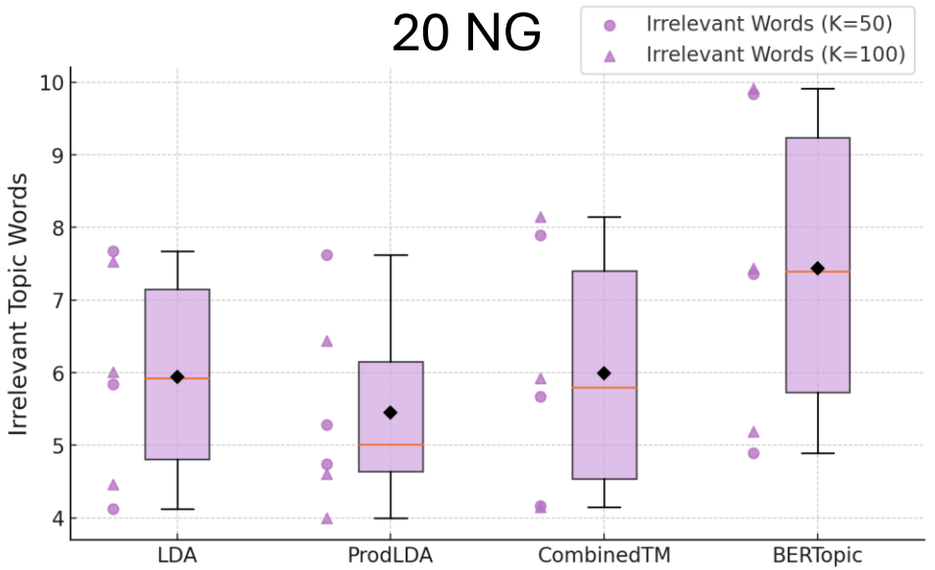
20 NG

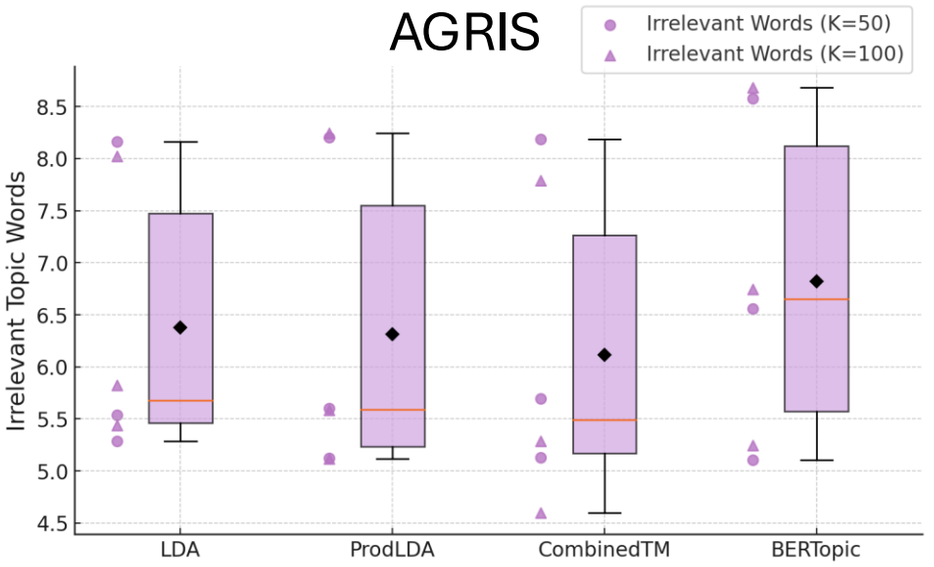


AGRIS



5. Results – Topic-document Alignment

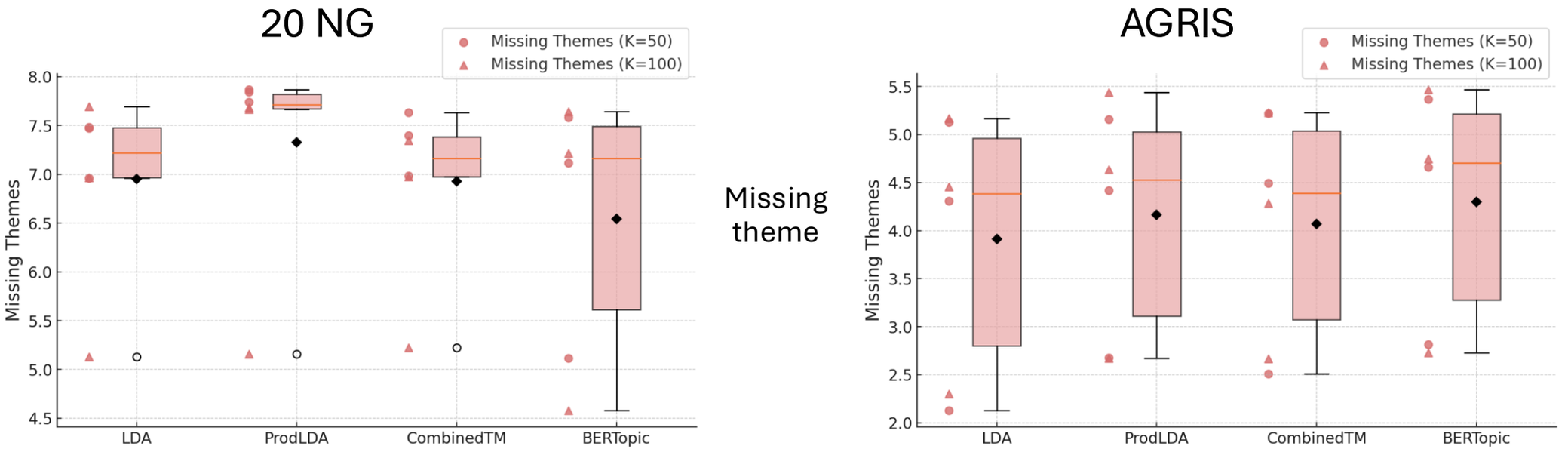




For detecting irrelevant topic words,

LDA, ProdLDA, and CombinedTM consistently yield lower counts compared to BERTopic, indicating that their topic words are more closely aligned with document content.

5. Results – Topic-document Alignment

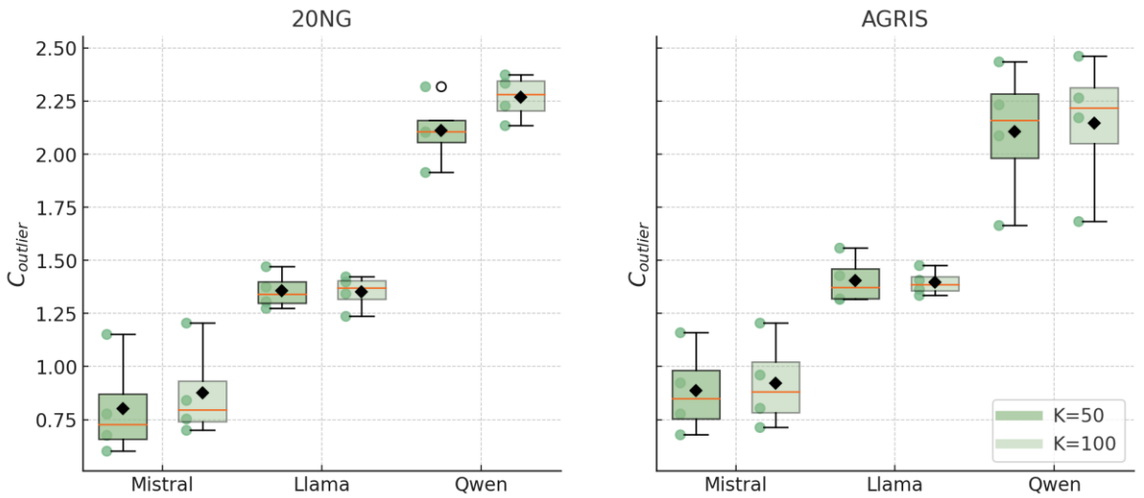


For detecting missing theme,

LDA, ProdLDA, and CombinedTM's performance are close to each other.
BERTopic missed the least themes on 20NG, while slightly higher than others on AGRIS.

6. Qualitative Analysis - Outlier Detection $C_{outlier}$

Topic words	Mistral	Llama	Qwen
interested, book, advance, fax, printer, print, email, address, mail, mailing	fax	fax, printer, print	advance, fax
keyboard, window, output, problem, work, time, run, input, response, drug	drug	drug	drug
science, evidence, theory, scientific, observation, scientist, fact, explain, bug, claim	bug	bug	bug, claim



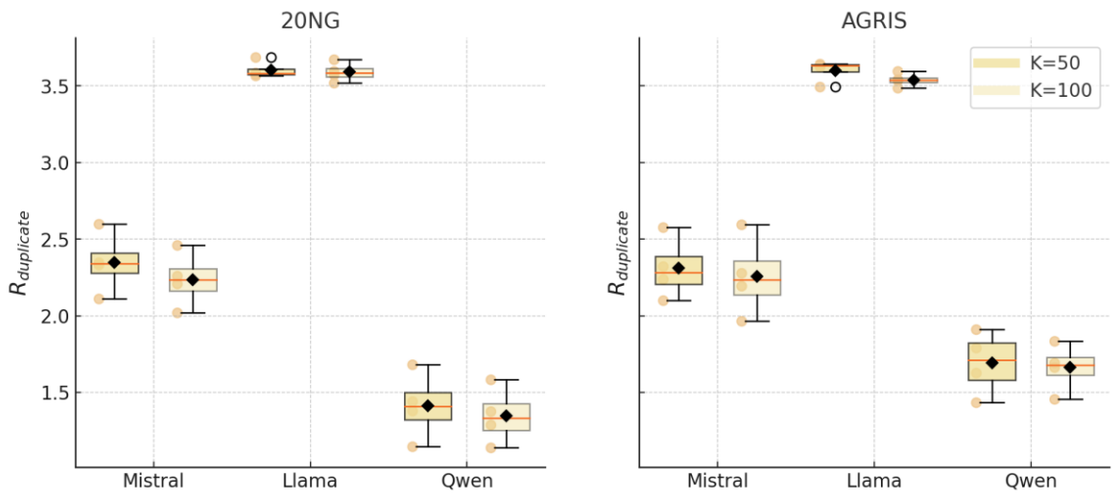
For detecting outliers in topic words,

Mistral is more cautious.

Qwen is relatively more aggressive in detecting words with unclear semantic pointing from topic words and considering them as outliers.

6. Qualitative Analysis - Duplicate Concept Detection $R_{duplicate}$

Topic words	Mistral	Llama	Qwen
faith, scripture, religion, moral, good, point, christian, church, belief, doctrine	(faith, belief), (christian, church)	(faith, religion), (christian, church), (doctrine, scripture)	(faith, belief), (scripture, doctrine)
disease, patient, health, medical, child, year, drug, treatment, adult, number	(patient, adult), (child, adult)	(disease, health), (patient, adult), (child, patient), (treatment, drug)	(disease, treatment), (medical, drug)
client, search, directory, package, software, file, mail, fax, database, project	(client, customer), (mail, email)	(client, directory), (search, package), (software, project)	None

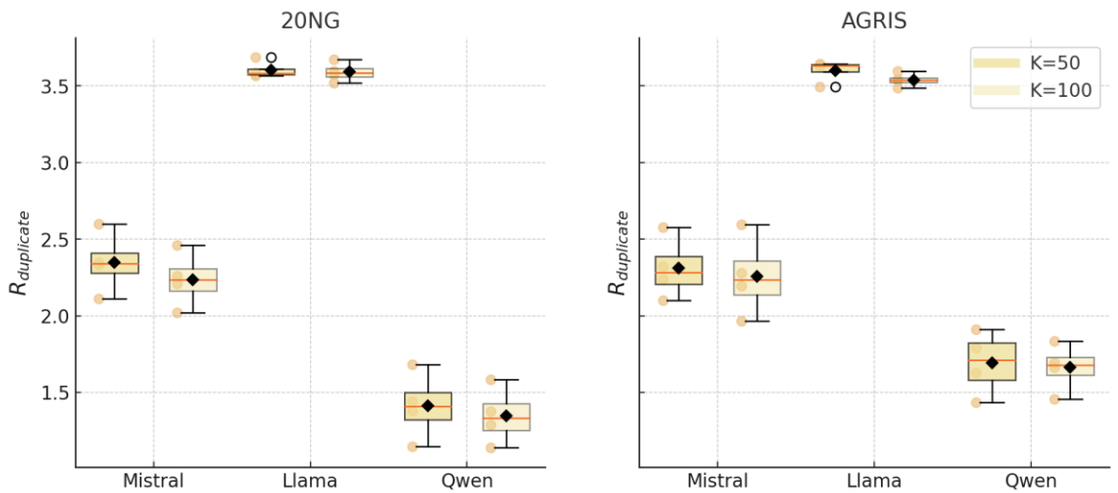


For detecting duplicate concept in topic words,

Mistral treats semantically related nouns (e.g., "christian" and "church", collective nouns where there is intersection (e.g., "patient" and "adult"), and nouns that belong to the same category (e.g. "child" and "adult") as conceptually identical.

6. Qualitative Analysis - Duplicate Concept Detection $R_{duplicate}$

Topic words	Mistral	Llama	Qwen
faith, scripture, religion, moral, good, point, christian, church, belief, doctrine	(faith, belief), (christian, church)	(faith, religion), (christian, church), (doctrine, scripture)	(faith, belief), (scripture, doctrine)
disease, patient, health, medical, child, year, drug, treatment, adult, number	(patient, adult), (child, adult)	(disease, health), (patient, adult), (child, patient), (treatment, drug)	(disease, treatment), (medical, drug)
client, search, directory, package, software, file, mail, fax, database, project	(client, customer), (mail, email)	(client, directory), (search, package), (software, project)	None

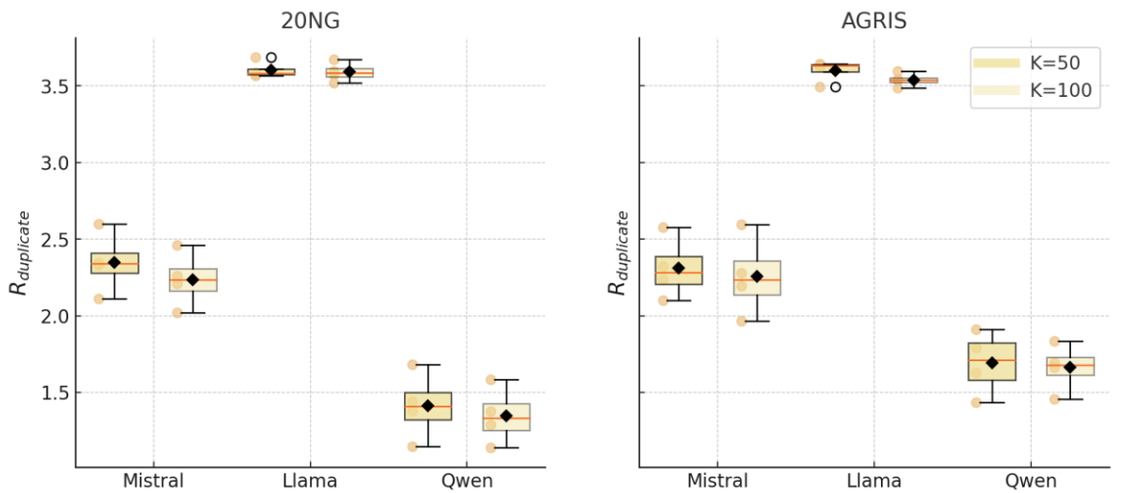


For detecting duplicate concept in topic words,

Also, Mistral had hallucinations (e.g., detecting a non-existent repetition of the word "customer" for "client" and a non-existent repetition of the word "email" for "mail").

6. Qualitative Analysis - Duplicate Concept Detection $R_{duplicate}$

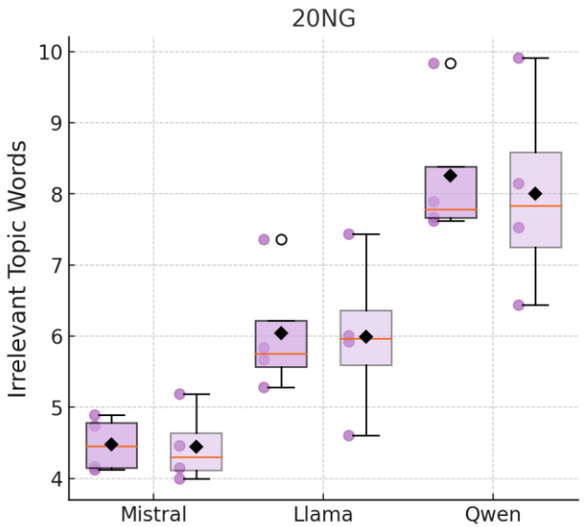
Topic words	Mistral	Llama	Qwen
faith, scripture, religion, moral, good, point, christian, church, belief, doctrine	(faith, belief), (christian, church)	(faith, religion), (christian, church), (doctrine, scripture)	(faith, belief), (scripture, doctrine)
disease, patient, health, medical, child, year, drug, treatment, adult, number	(patient, adult), (child, adult)	(disease, health), (patient, adult), (child, patient), (treatment, drug)	(disease, treatment), (medical, drug)
client, search, directory, package, software, file, mail, fax, database, project	(client, customer), (mail, email)	(client, directory), (search, package), (software, project)	None

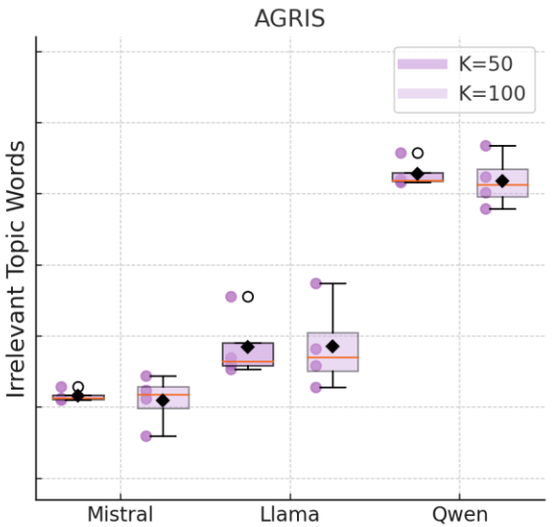


For detecting duplicate concept in topic words,

Llama treats grammatically related words (e.g., the verb “search” and its potential object “ package”), semantically opposite words (e.g., “disease” and “health “) as conceptually identical.

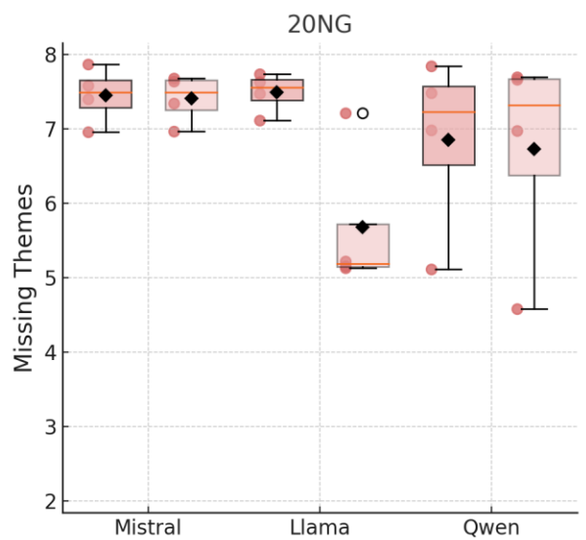
6. Qualitative Analysis

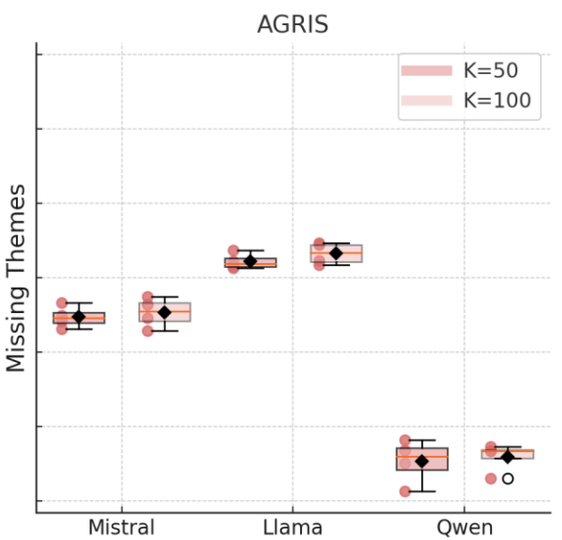




For detecting irrelevant topic words,

Qwen tend to detect lot more irrelevant topic words than Mistral. Llama is in a middle position.





For detecting missing theme,

One interesting facts is, all LLMs found lot more missing theme on dataset 20NG than on AGRIS.

Conclusion

- Introduced a comprehensive framework for evaluating topic models using LLM-based metrics that complements traditional automated metrics by incorporating nuanced measures of coherence, repetitiveness, diversity, and topic–document alignment.
- Reveal not only the strengths and weaknesses of various topic models,

  but also the intrinsic biases and judgment tendencies of different LLM evaluators.

Any comments or question?

Email: zhiyin.tan@l3s.de

Github: https://github.com/zhiyintan/topic-model-LLMjudgment