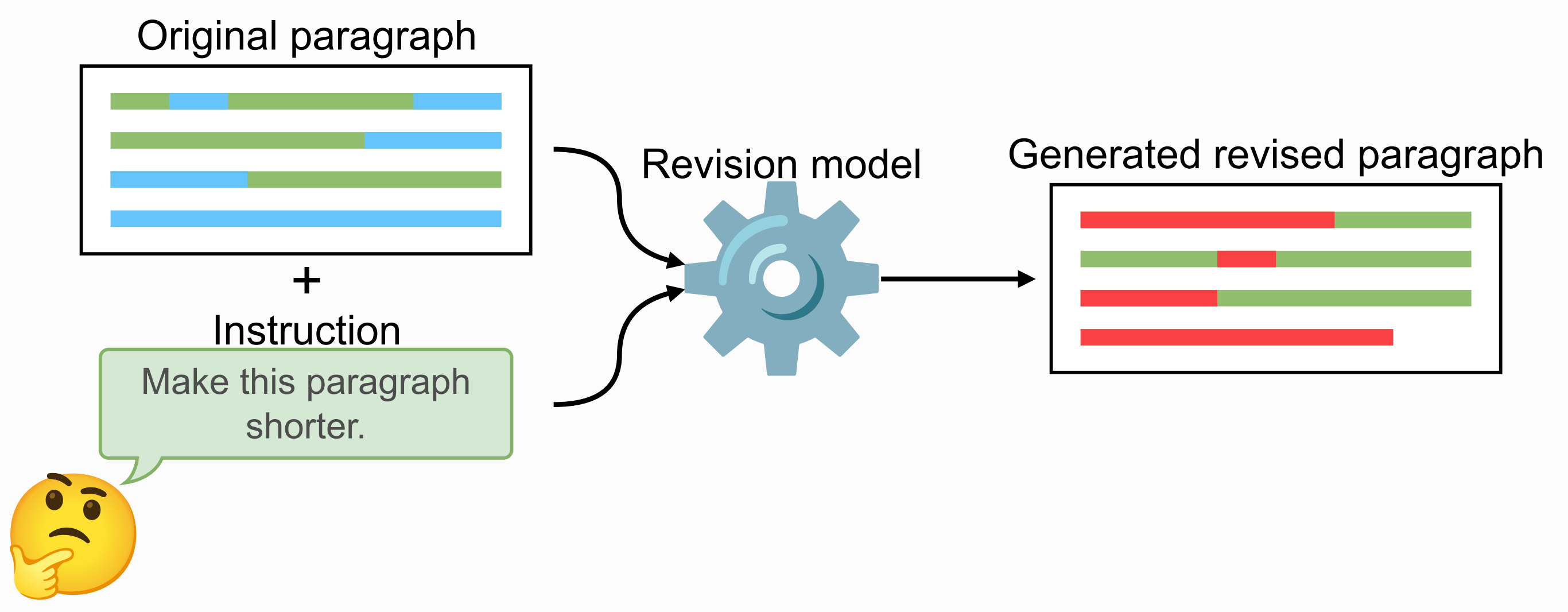
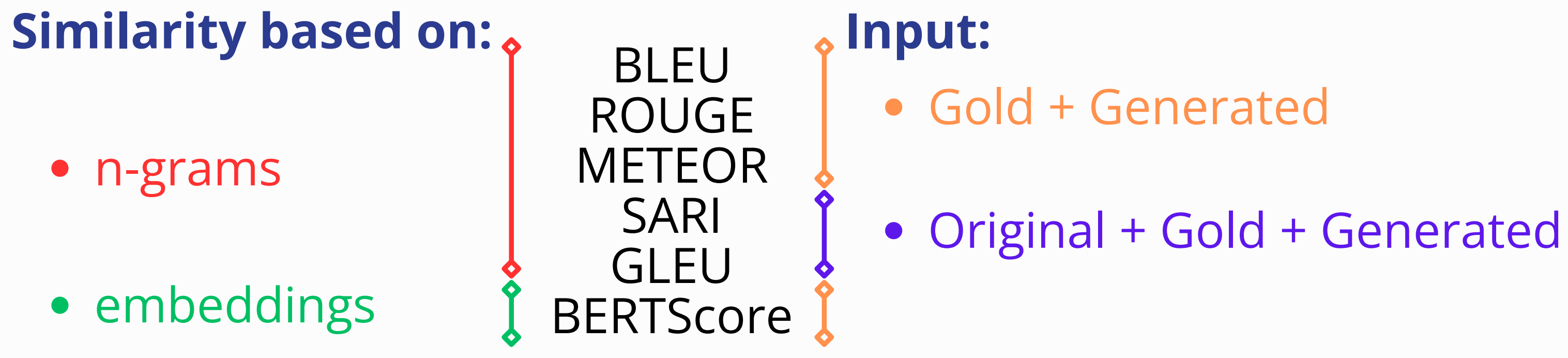
Identifying Reliable Evaluation Metrics for Scientific Text Revision

Léane Jourdan[♣], Florian Boudin[♦], Richard Dufour[♣], Nicolas Hernandez[♣]
{firstname.lastname}@univ-nantes.fr
♣ LS2N, Nantes Université ♦ JFLI, NII, Tokyo

Paragraph Revision Task



Usually evaluated using


Similarity based on:

- n-grams
- embeddings

BLEU
ROUGE
METEOR
SARI
GLEU
BERTScore

Input:

- Gold + Generated
- Original + Gold + Generated

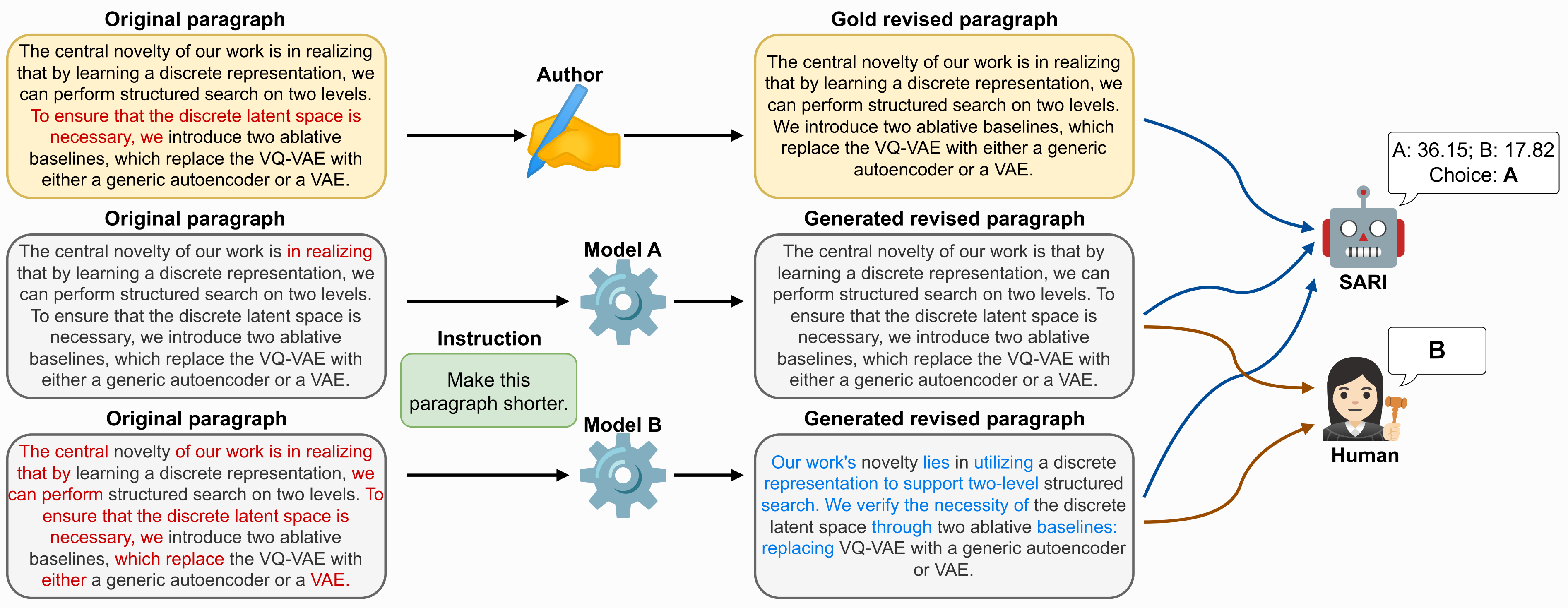
Initial experiment

Dataset
***ParaRev** test set*

- 258 pairs of revised paragraphs
- 2 annotations/paragraph
  - =516 datapoints

We evaluated 6 revision models with the similarity metrics

- no edits baseline achieve the highest scores
- Approach conducting minimal edits also achieve strong scores


Original paragraph
The central novelty of our work is in realizing that by learning a discrete representation, we can perform structured search on two levels. To ensure that the discrete latent space is necessary, we introduce two ablative baselines, which replace the VQ-VAE with either a generic autoencoder or a VAE.

Gold revised paragraph
The central novelty of our work is in realizing that by learning a discrete representation, we can perform structured search on two levels. We introduce two ablative baselines, which replace the VQ-VAE with either a generic autoencoder or a VAE.

Original paragraph
The central novelty of our work is in realizing that by learning a discrete representation, we can perform structured search on two levels. To ensure that the discrete latent space is necessary, we introduce two ablative baselines, which replace the VQ-VAE with either a generic autoencoder or a VAE.

Generated revised paragraph
The central novelty of our work is that by learning a discrete representation, we can perform structured search on two levels. To ensure that the discrete latent space is necessary, we introduce two ablative baselines, which replace the VQ-VAE with either a generic autoencoder or a VAE.

Original paragraph
The central novelty of our work is in realizing that by learning a discrete representation, we can perform structured search on two levels. To ensure that the discrete latent space is necessary, we introduce two ablative baselines, which replace the VQ-VAE with either a generic autoencoder or a VAE.

Generated revised paragraph
Our work's novelty lies in utilizing a discrete representation to support two-level structured search. We verify the necessity of the discrete latent space through two ablative baselines: replacing VQ-VAE with a generic autoencoder or VAE.

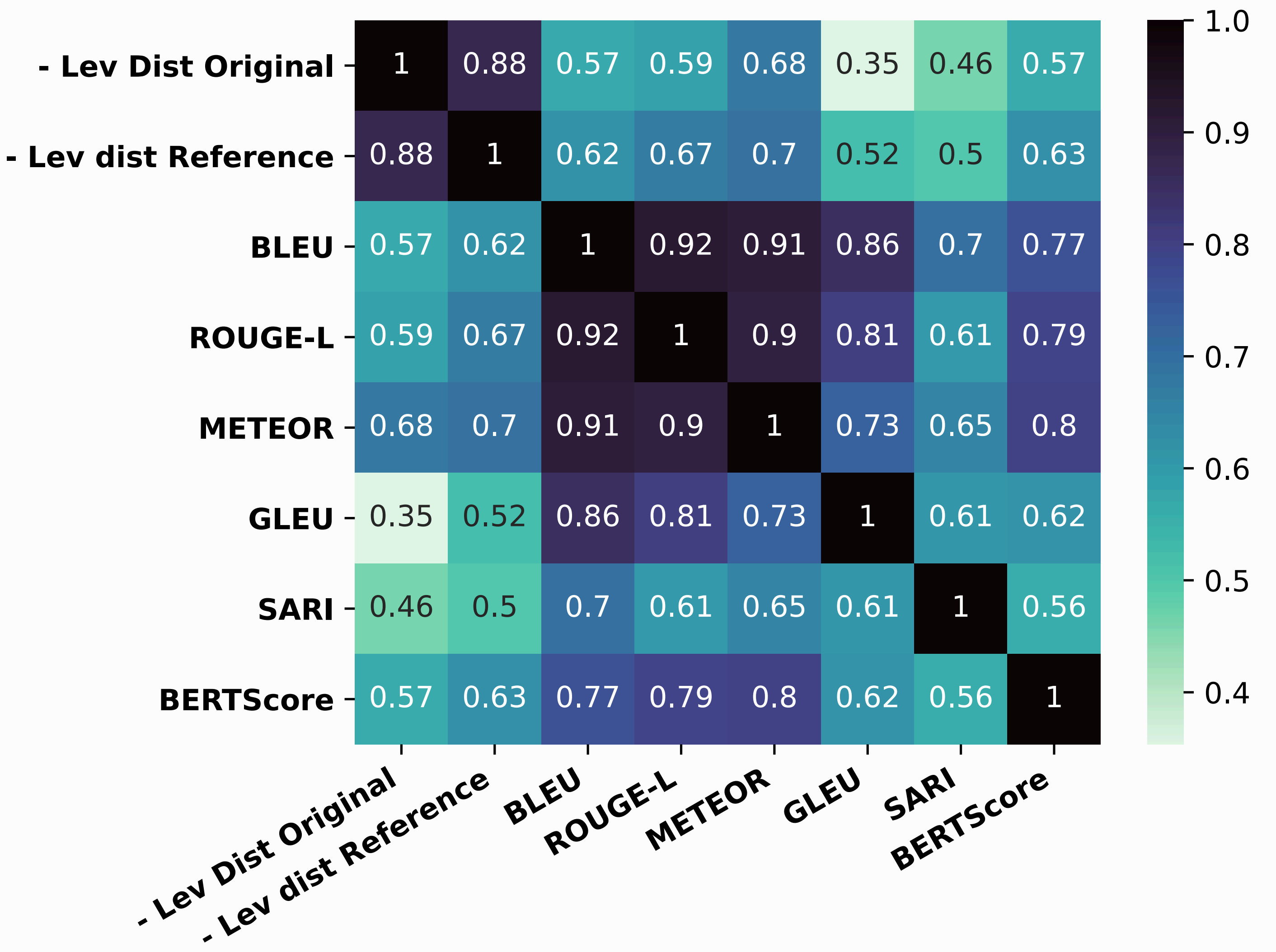
Instruction
Make this paragraph shorter.

Limitations of Similarity-based Metrics

1. **Redundancy and Correlation Among Metrics**
   Most metrics are highly correlated, providing **redundant information**.
2. **Metrics only capture surface similarity**
   Reflect how closely a model **replicates** the reference revision, **rather than** evaluating the **quality** of the revision itself.
3. **Substantial revisions are penalised**
   The more a revision deviates from the original paragraph, the lower its score.
   The metrics **do not reward** substantial, **qualitative improvements**.

Resulting in evaluation bias:
Making no revision at all or minimal edits will often result in a higher score than making meaningful changes.

SARI and GLEU stand out as **exceptions**, and are the only ones to consider the source.



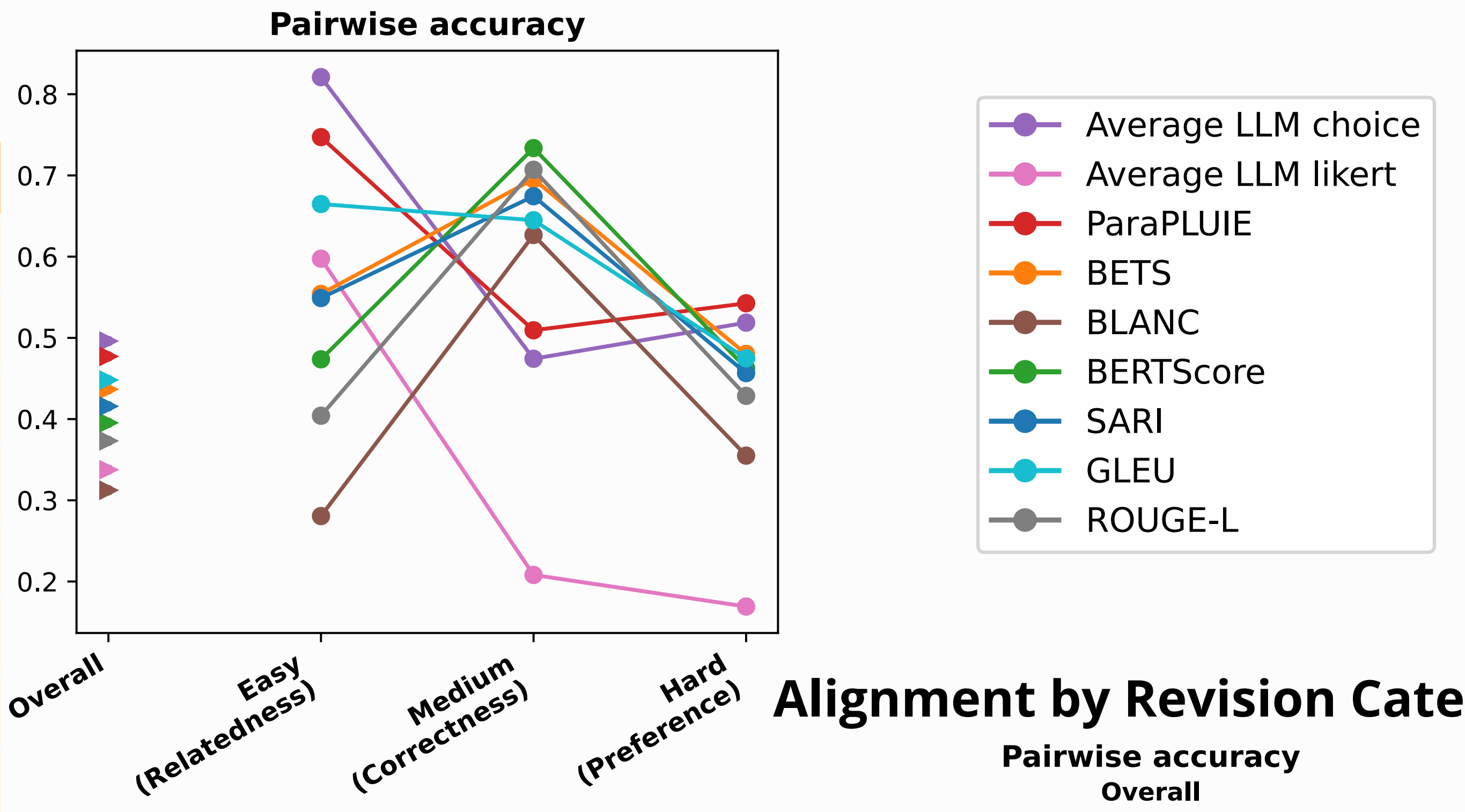
	- Lev Dist Original	- Lev dist Reference	BLEU	ROUGE-L	METEOR	GLEU	SARI	BERTScore
- Lev Dist Original	1	0.88	0.57	0.59	0.68	0.35	0.46	0.57
- Lev dist Reference	0.88	1	0.62	0.67	0.7	0.52	0.5	0.63
BLEU	0.57	0.62	1	0.92	0.91	0.86	0.7	0.77
ROUGE-L	0.59	0.67	0.92	1	0.9	0.81	0.61	0.79
METEOR	0.68	0.7	0.91	0.9	1	0.73	0.65	0.8
GLEU	0.35	0.52	0.86	0.81	0.73	1	0.61	0.62
SARI	0.46	0.5	0.7	0.61	0.65	0.61	1	0.56
BERTScore	0.57	0.63	0.77	0.79	0.8	0.62	0.56	1

Results

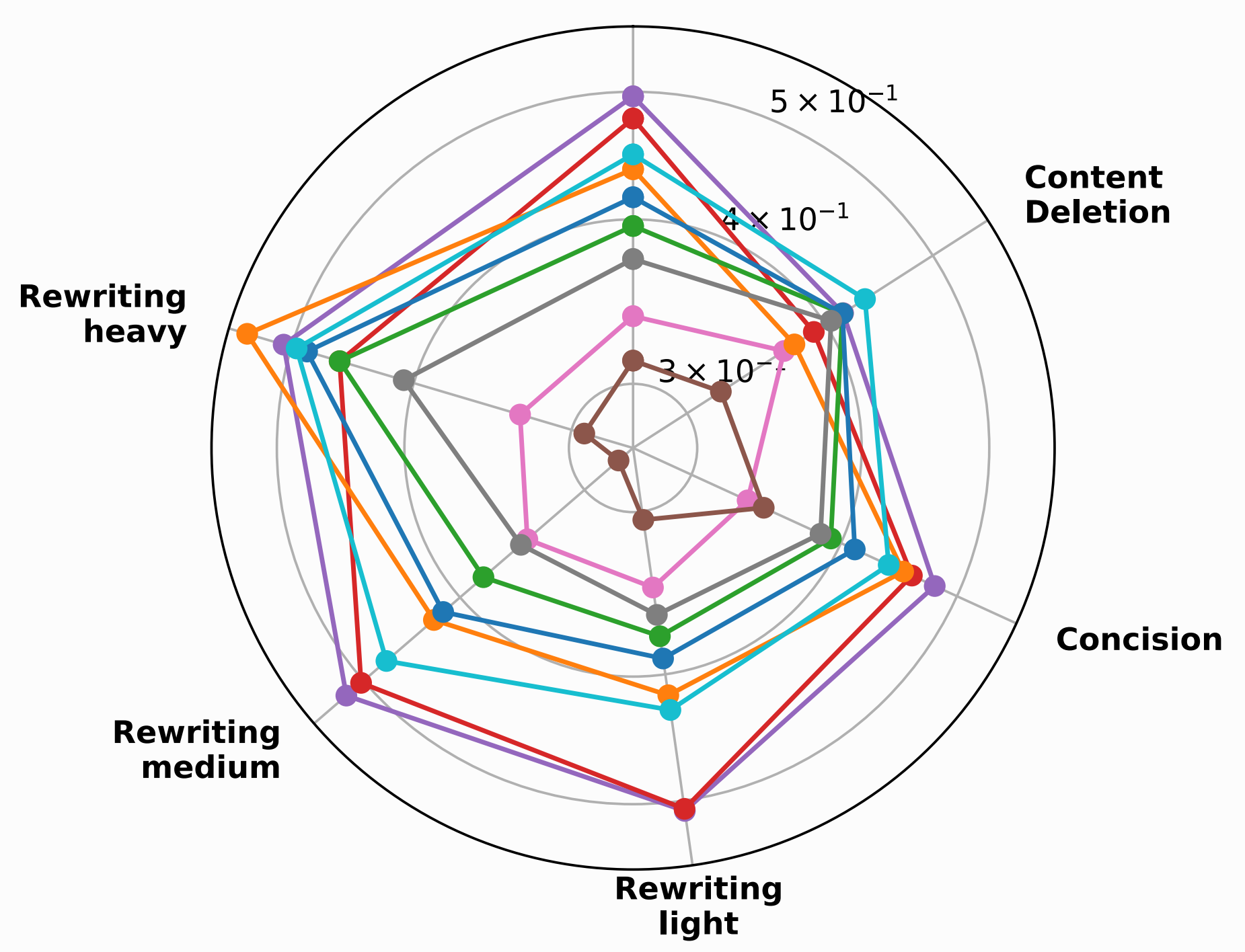
Alignment of automatic metrics with human judgements

Judge	Pair acc.	V	κ
Avg. LLM choice	0.496	0.239	0.247
Avg. LLM likert	0.338	0.240	0.181
ParaPLUIE	0.477	0.225	0.197
BETS	0.437	0.152	0.127
BLANC	0.312	0.117	-0.080
Bertscore	0.395	0.161	0.034
SARI	0.416	0.184	0.071
GLEU	0.448	0.193	0.138
ROUGE-L	0.373	0.179	-0.013
Random	*0.264*	*0.027*	*-0.006*

Alignment by Difficulty



Alignment by Revision Category



Proposed Alternative Evaluation

1. **Instruction following**: LLM-as-a-judge
2. **Similarity to gold**: SARI or GLEU
3. **Meaning-preservation**: ParaPLUIE

Candidate Metrics

From Related NLP Domains
Input: Original + Generated
BETS: Text simplification
BLANC: Summarization
ParaPLUIE: Paraphrase detection

LLM-as-a-judge
Input: Original + Generated (+ Gold)
Based on the 3 manual evaluation criteria
LLM-choice (Yes/No + Pairwise comparison)
LLM-Likert (Generating Scores)

Findings

- **ParaReval, a dataset** of human pairwise evaluations of generated revisions
- Traditional similarity metrics alone **fail to accurately evaluate** text revision
- An **alternative evaluation method** composed of 3 complementary metrics

ACL 2025 VIENNA
Nantes Université
TALN
LS2N Laboratoire des Sciences du Numérique de Nantes