

Sarah Frank^{1,2}, Sebastian Schäffer², Andreas Wagner¹, Christian Gütl²

CREATING EXPLAINABLE SUMMARIES FOR LONG SCIENTIFIC DOCUMENTS USING LARGE LANGUAGE MODELS

¹CERN, Switzerland

²Graz University of Technology, Austria



PROBLEM

Efficiently acquire the most significant information from scientific papers
for applications and use cases related to science search

RQ01

How do authors evaluate the quality of the
summaries generated by the AI system
according to selected metrics?

RQ02

Can this approach improve the
transparency of the summary and does
this influence user trust?

IDEA

Create a system that:

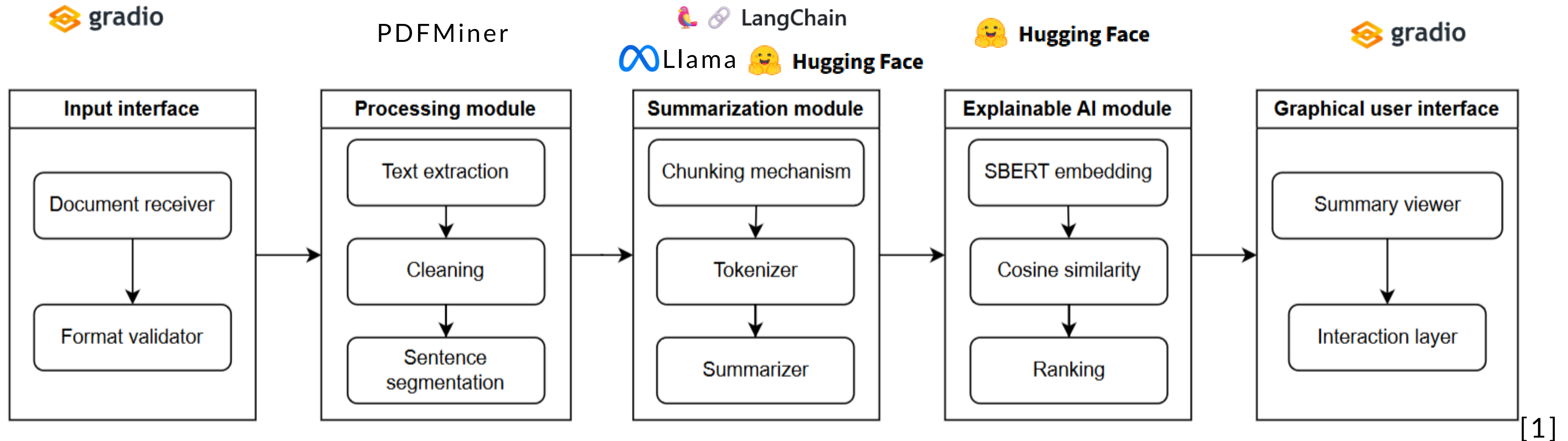
- Summarizes selected scientific papers from PDF
- Allows for users to see where information was sourced in the paper
- Displays this information in a way that furthers trustability

SUMMARIZATION

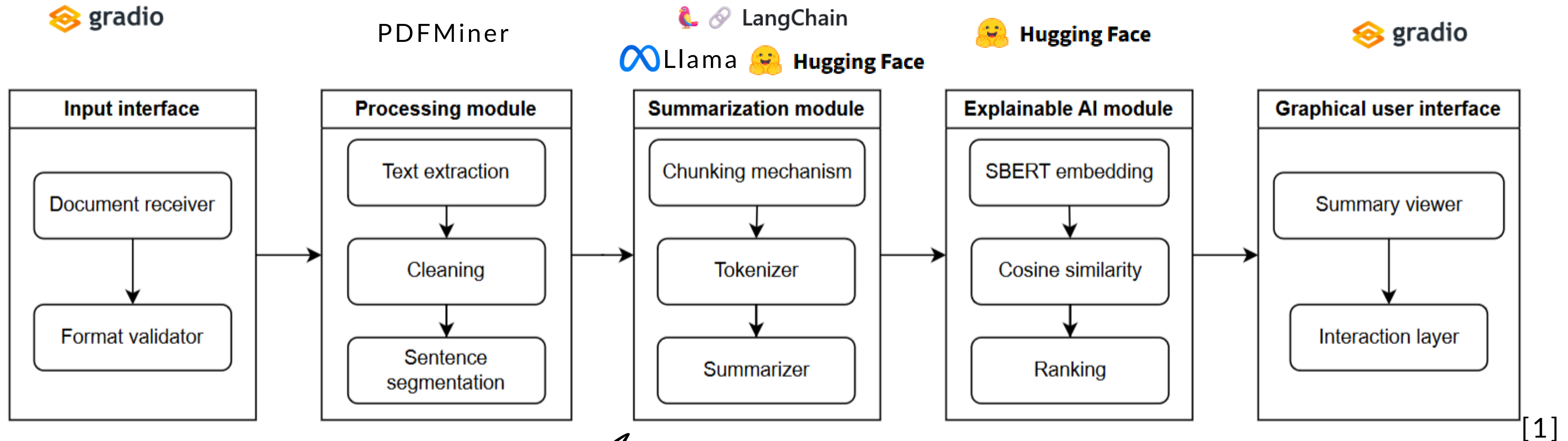
EXPLAINABILITY

VISUALIZATION

CONCEPT & IMPLEMENTATION



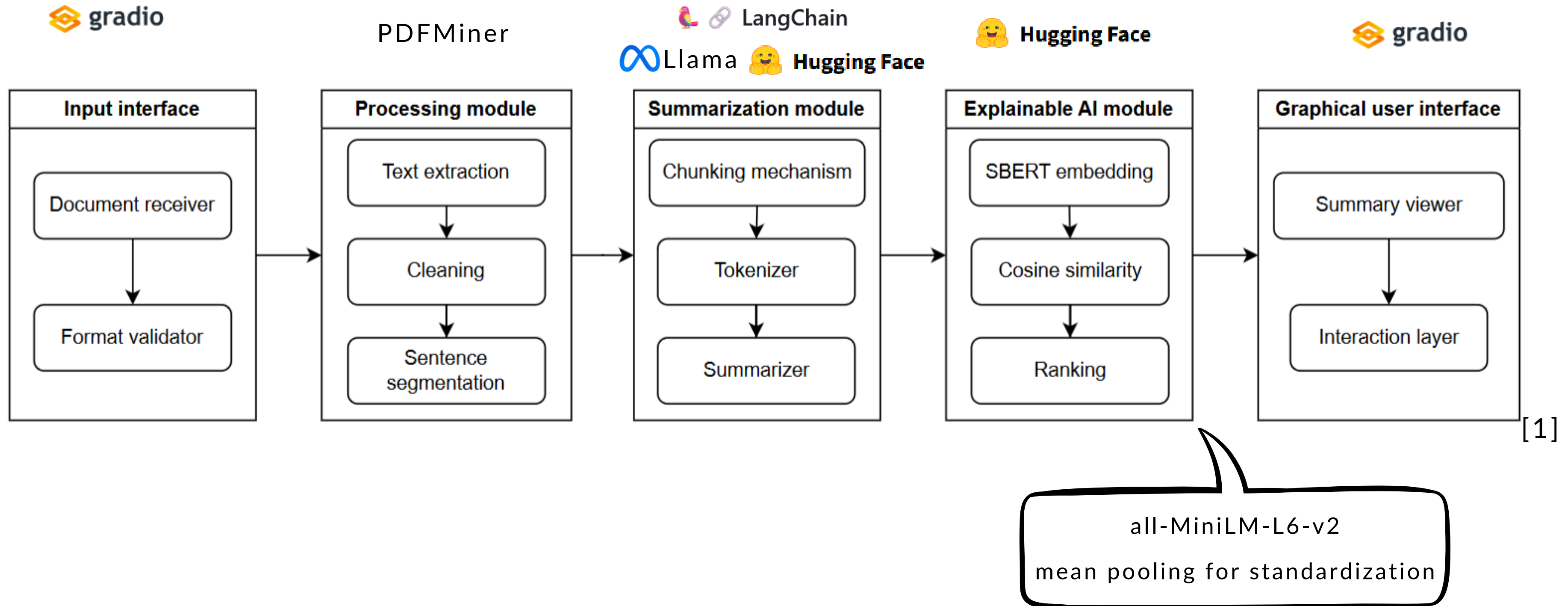
CONCEPT & IMPLEMENTATION



Llama-2-7b-Chat: 4096 tokens
50 character chunk overlap


[1] From "Summarizing Long Scientific Documents: Leveraging Llama2-7B-Chat with Explainable AI" by S. Schäffer, p. 53. Adapted with permission

CONCEPT & IMPLEMENTATION



[1] From "Summarizing Long Scientific Documents: Leveraging Llama2-7B-Chat with Explainable AI" by S. Schäffer, p. 53. Adapted with permission

Upload a PDF file



Drop File Here
- or -
Click to Upload

Select a sentence from the summary

Submit

Most similar sentence in the original document

Summary

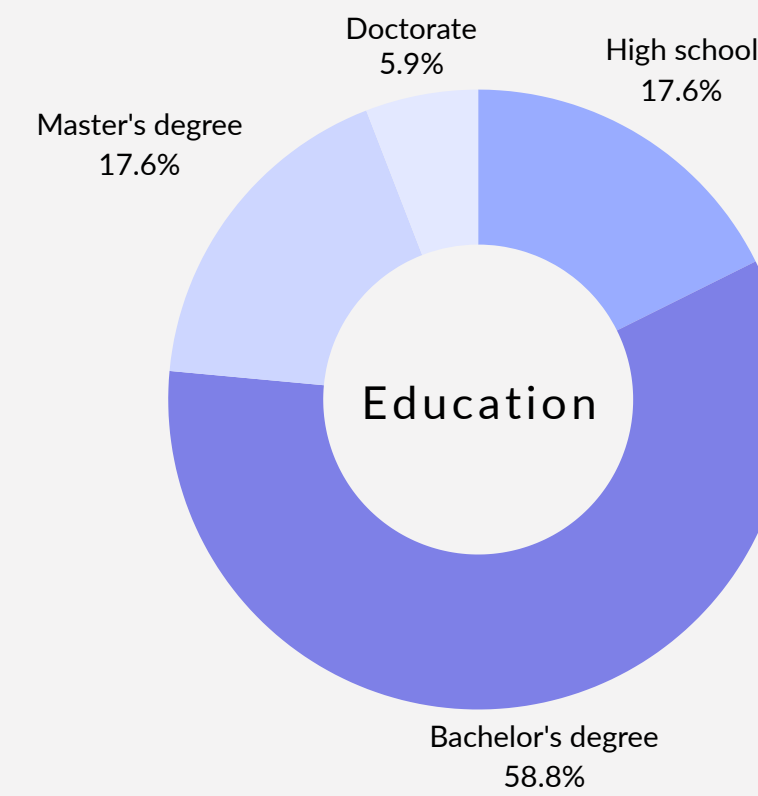
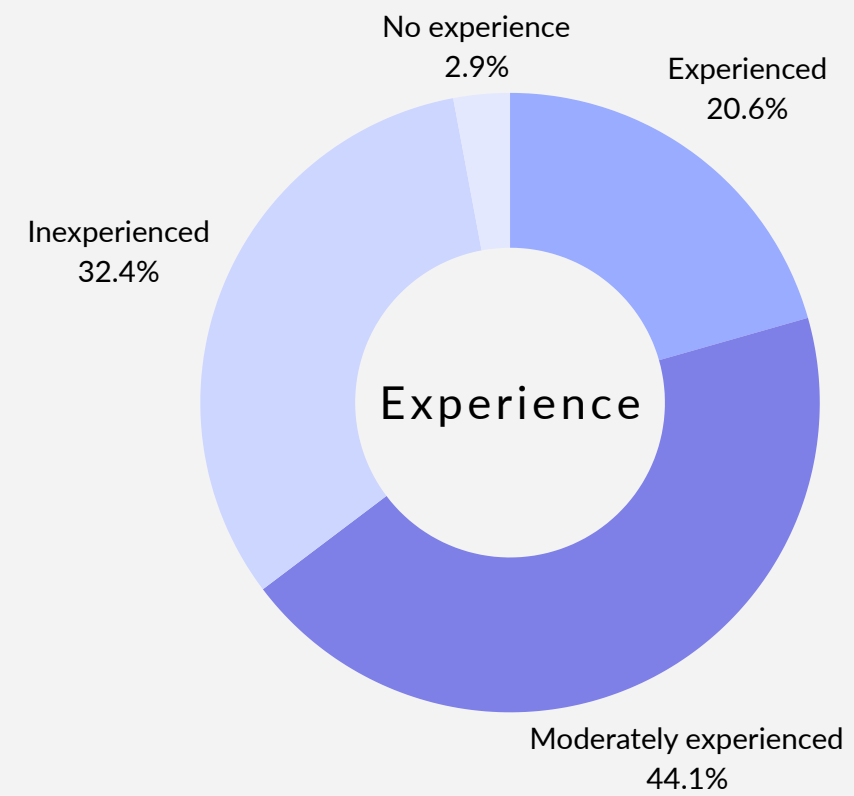
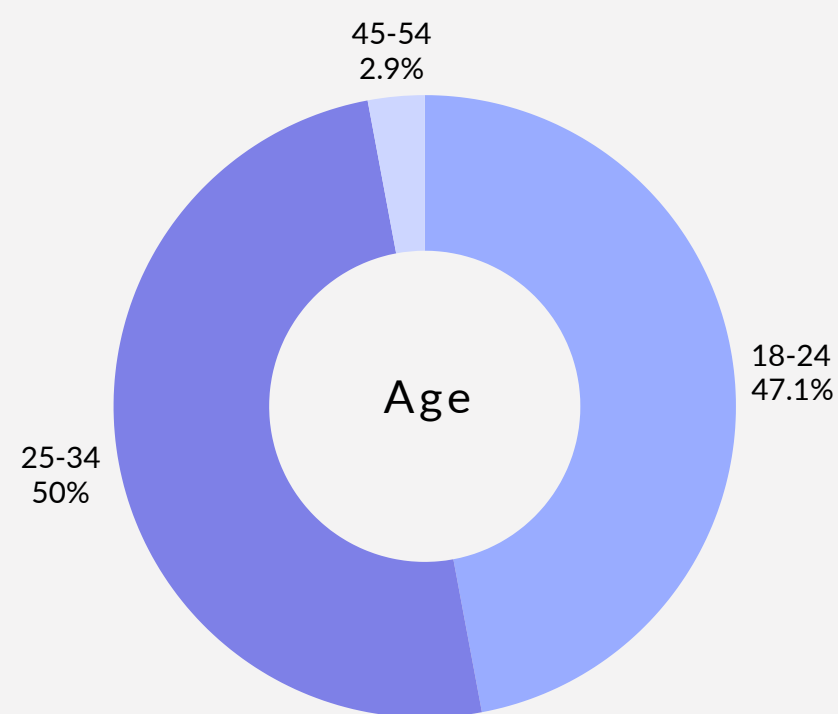
AUTHOR STUDY

- 20 summaries for papers published in the Journal of University Computer Science
- 68 authors invited, 11 participants

Metric	Very low	Low	Moderate	High	Very high	Mean	Stdev
Accuracy	1	1	0	6	3	3.82	1.25
Key contributions	1	1	2	3	4	3.73	1.35
Coherence	1	1	4	4	1	3.27	1.10
Overall satisfaction	1	0	3	5	2	3.64	1.12

EXPLAINABILITY STUDY

- Open study
- 34 participants



EXPLAINABILITY STUDY

- Increased time efficiency
- Transparency may not be sufficient or the quality of summaries insufficient

Metric	Very low	Low	Moderate	High	Very high	Mean	Stdev
Clarity	0	1	5	21	7	4.00	0.69
Trust in accuracy	1	3	15	11	4	3.41	0.91
Explainability	0	2	7	13	12	4.03	0.89
Coherence	0	2	4	17	11	4.09	0.82
Effect on trust	0	3	4	20	7	3.91	0.82
Interaction support	0	1	7	18	8	3.97	0.75



DISCUSSION

- Coherence scores differ significantly between the two groups
- Significant standard deviation across all metrics in author survey
- Explainability module rated “high” for influence on trust by 58.82% of survey participants
- Low trust in accuracy with high standard deviation
- Points of criticism:
 - Summary length
 - Redundancy
 - Text cohesion
 - Monotonous
 - Reliability

CONCLUSION

- Trustable summarization system for scientific articles
- Author and general user perspectives
- Promising results but high variation
- Summary quality and transparency-increasing measures to be improved

THANK YOU

Sarah Frank
sarah.frank@cern.ch