Annotation Quality in Aspect-Based Sentiment Analysis: A Case Study Comparing Experts, Students, Crowdworkers, and Large Language Models

Niklas Donhauser
Media Informatics Group
University of Regensburg

Jakob Fehle
Media Informatics Group
University of Regensburg

Nils Constantin Hellwig
Media Informatics Group
University of Regensburg

Markus Weinberger
Media Informatics Group
University of Regensburg

Christian Wolff
Media Informatics Group
University of Regensburg

Udo Kruschwitz
Information Science Group
University of Regensburg

Universität Regensburg

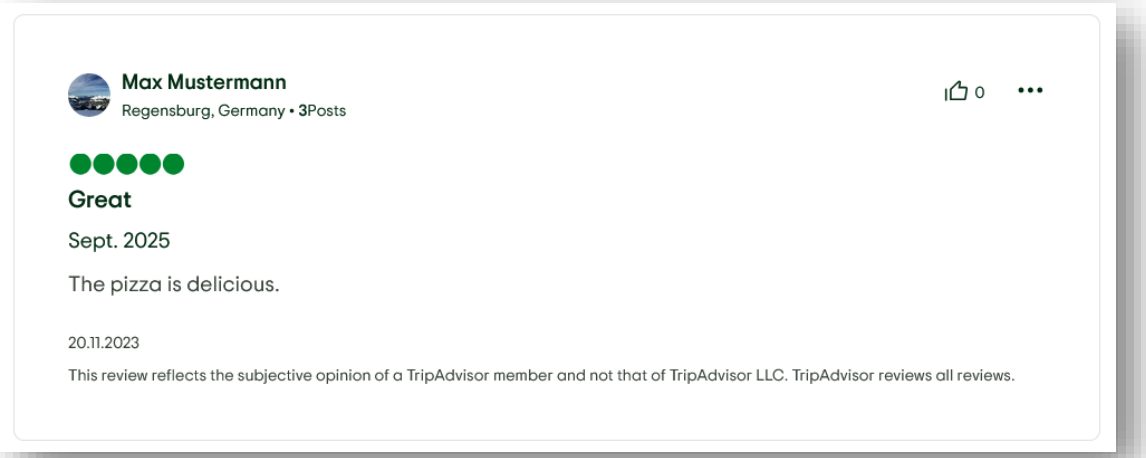
LREC 2026 Palma

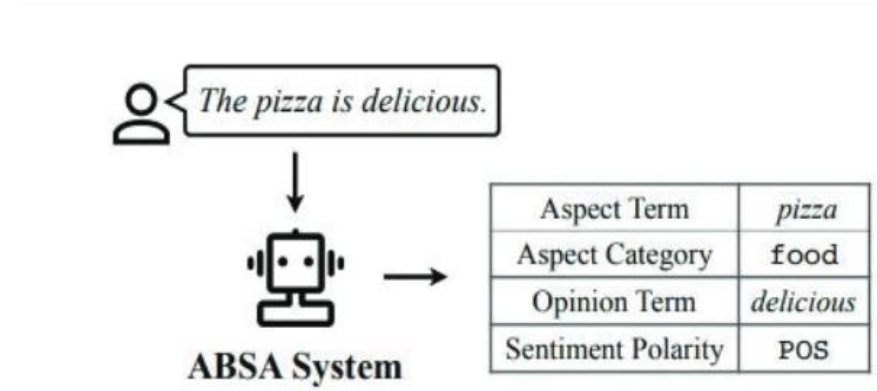
Introduction

Background

- **Sentiment Analysis (SA):**
  Evaluation of the general sentiment (positive / negative / neutral) in texts [1]

- **Aspect-based Sentiment Analysis (ABSA):**
  Analysis of the sentiment towards certain aspects of an entity (e.g. properties, product features) [2]





Sentiment Elements [3]

[1] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature. https://hyse.org/pdf/SentimentAnalysis-and-OpinionMining.pdf
[2] Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review, 49, 100576. https://doi.org/10.1016/j.cosrev.2023.100576
[3] Singhi, V., Chauhan, C., & Soni, P. K. (2024, April). Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-10). IEEE. https://doi.org/10.1109/I2CT61223.2024.10543612

! Motivation

- Strong progress in English, but German lacks high-quality datasets
- Dataset creation is costly & depends on annotation quality
- Unclear: How do different annotators (experts, crowdworkers, students, LLMs) affect model performance?

➢ **What is the impact of annotation quality and different annotator groups on data and model quality for ABSA?**

Research Goals

- Compare annotation quality across experts, students, crowdworkers, and LLMs
- Measure impact on inter-annotator agreement (IAA) and dataset reliability
- Evaluate downstream performance on ABSA tasks

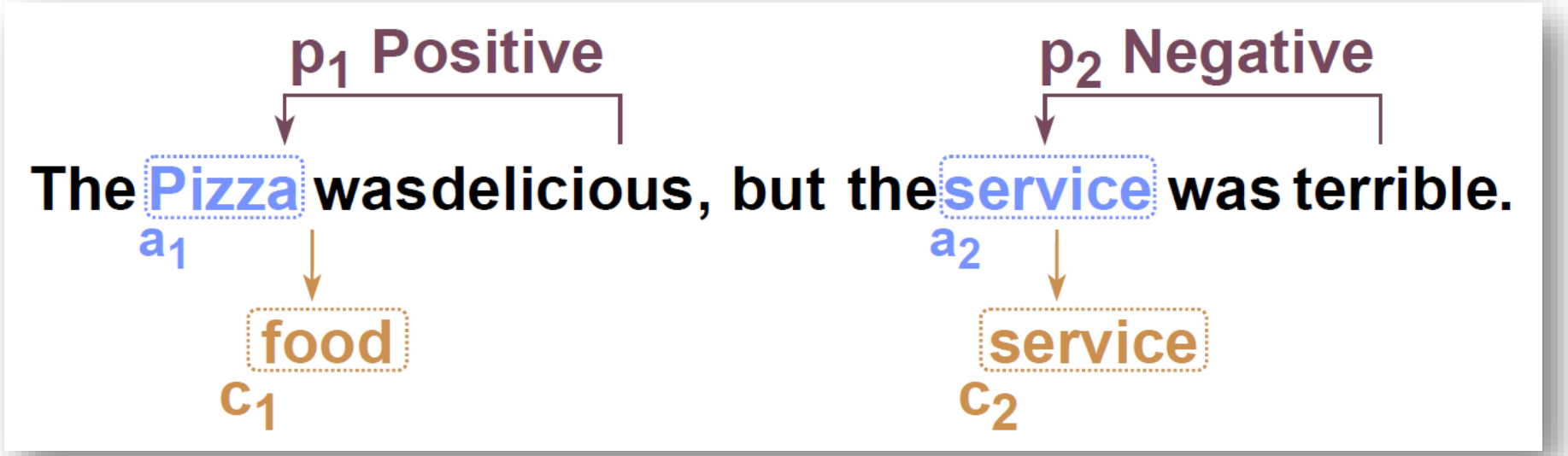
Annotation Objective

Aspect Category Sentiment Analysis (ACSA)

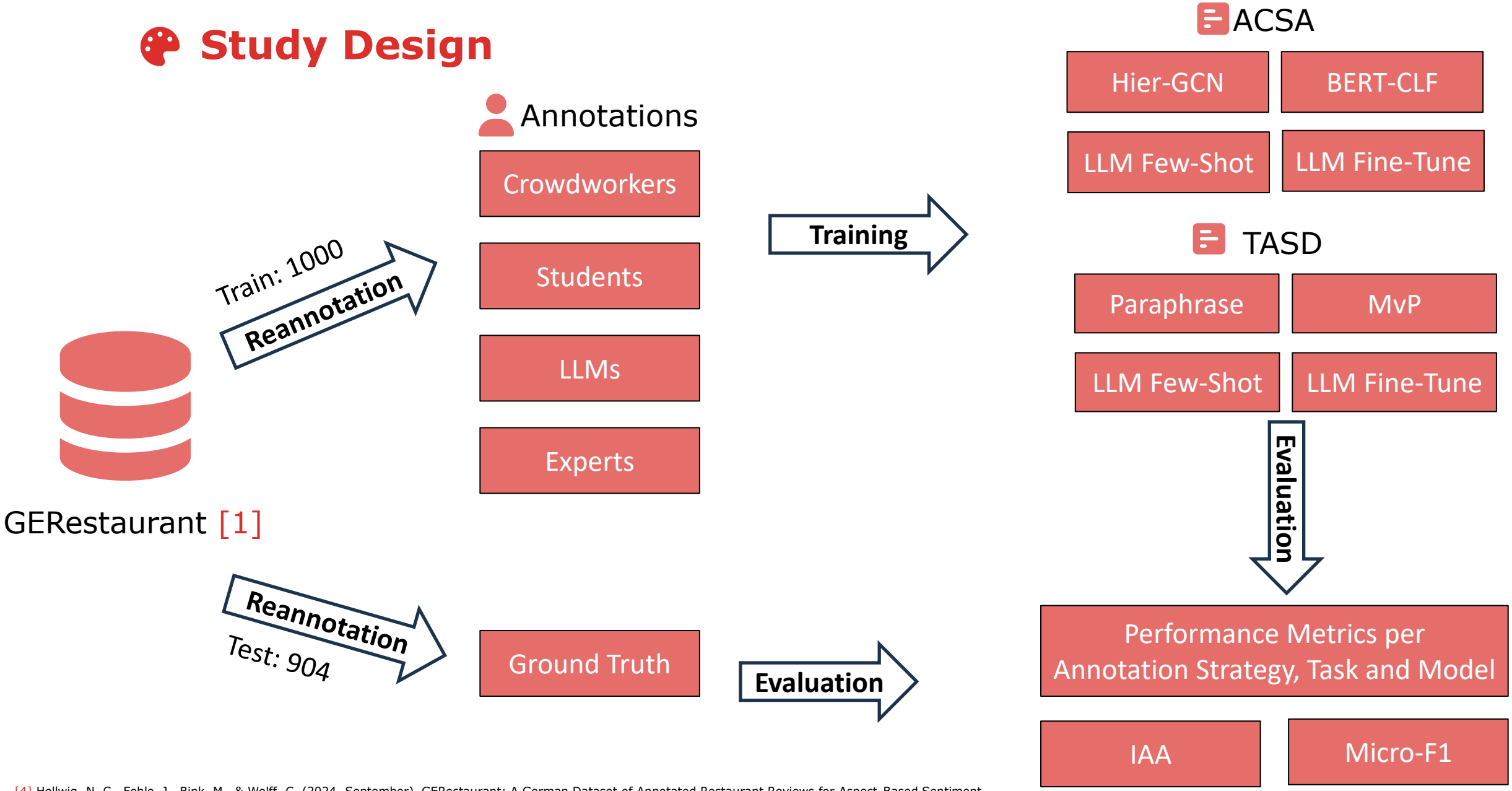
- Aspect Category: food (c_1), service (c_2)
- Aspect Polarity: positive (p_1), negative (p_2)

Target Aspect Sentiment Detection (TASD)

- Aspect Category: food (c_1), service (c_2)
- Aspect Polarity: positive (p_1), negative (p_2)
- Aspect Term: Pizza (a_1), service (a_2)



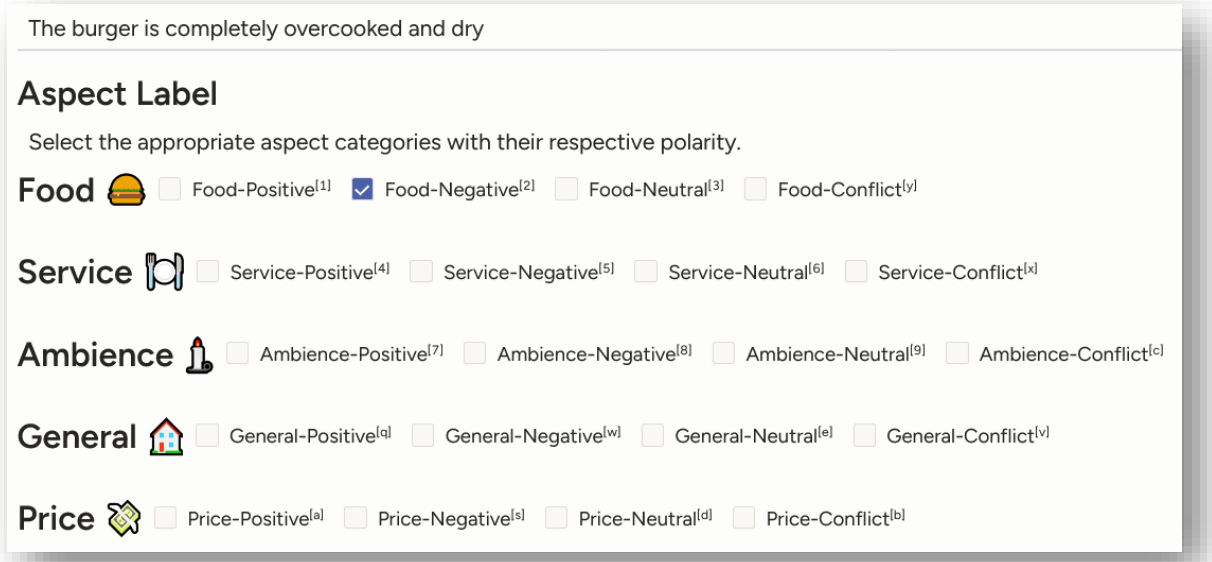
Study Design



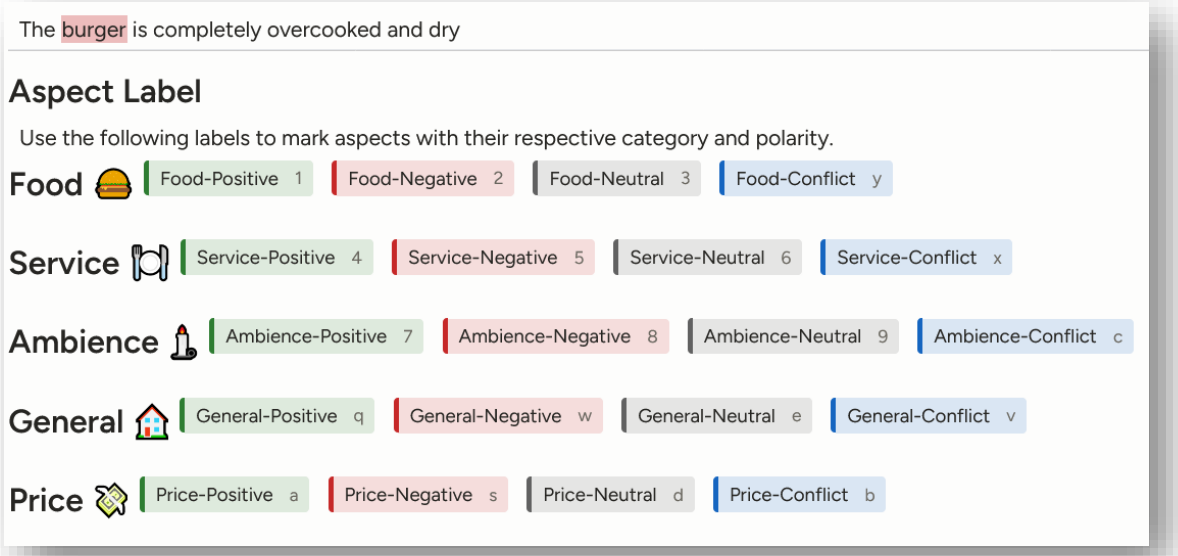
[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

Annotation tool

- Label Studio[1] as the annotation tool
- New annotation of 1000 random sampled data entries from the original train set



(a) Label interface for the ACSA task in Label Studio.

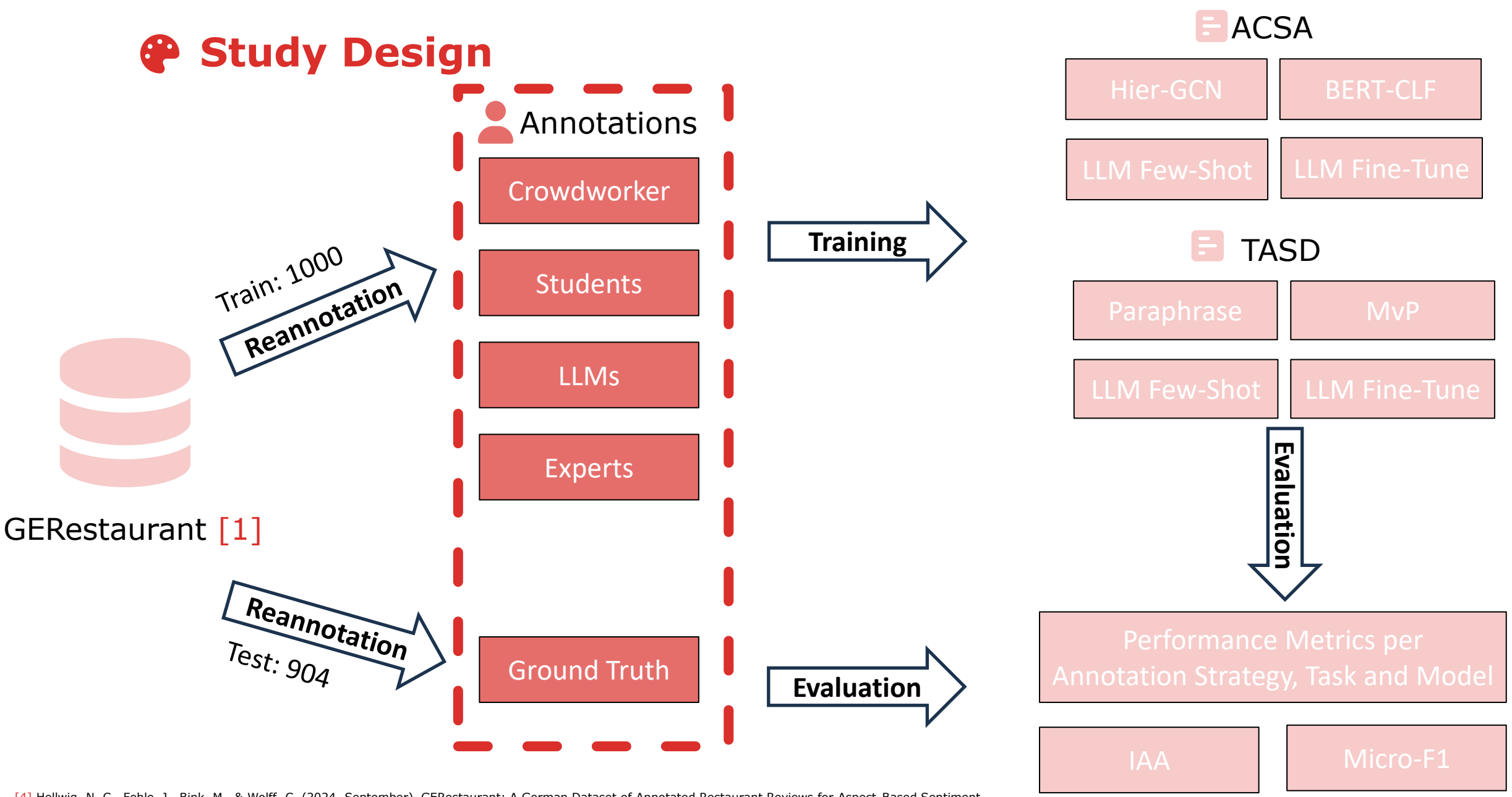


(b) Label interface for the TASD task in Label Studio.

[1] Label Studio: https://labelstudio.io/

Annotation

Study Design



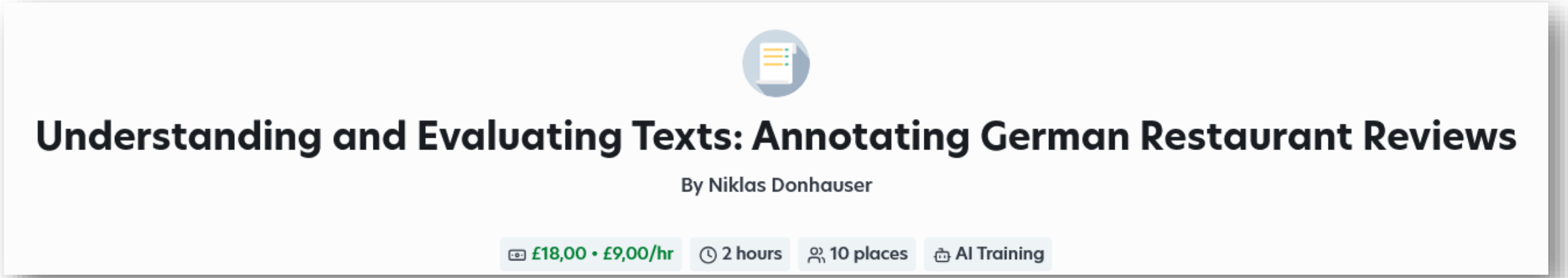
[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

Annotation: Ground Truth

- 924 sentences annotated independently by two ABSA-experienced annotators
- Iterative guideline refinement based on disagreement patterns
- Joint decision-making in the case of disagreement

Annotation: Crowdworkers

- 30 German-speaking participants recruited via Prolific[1] (DE/AT/CH)
- Paid £9/hour; quality controlled via Prolific filters and timing checks
- Pre-study questionnaire and informed consent collected via Google Forms



[1] Prolific: https://www.prolific.com/

Annotation: Large Language Models

- Based on [5] using Gemma 3 27B[1] for dataset annotation
- 30 few-shot examples used per prompt (cost-efficient setup)
- Outputs aggregated via majority voting across seeds (self-consistency)
- Final annotations included only if supported by majority of runs

[5] Hellwig, N. C., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, May). *Do we still need human annotators? Prompting large language models for aspect sentiment quad prediction*. arXiv. https://doi.org/10.48550/arXiv.2502.13044
[3]Gemma 3 27B https://huggingface.co/google/gemma-3-27b-it

Annotation: Students

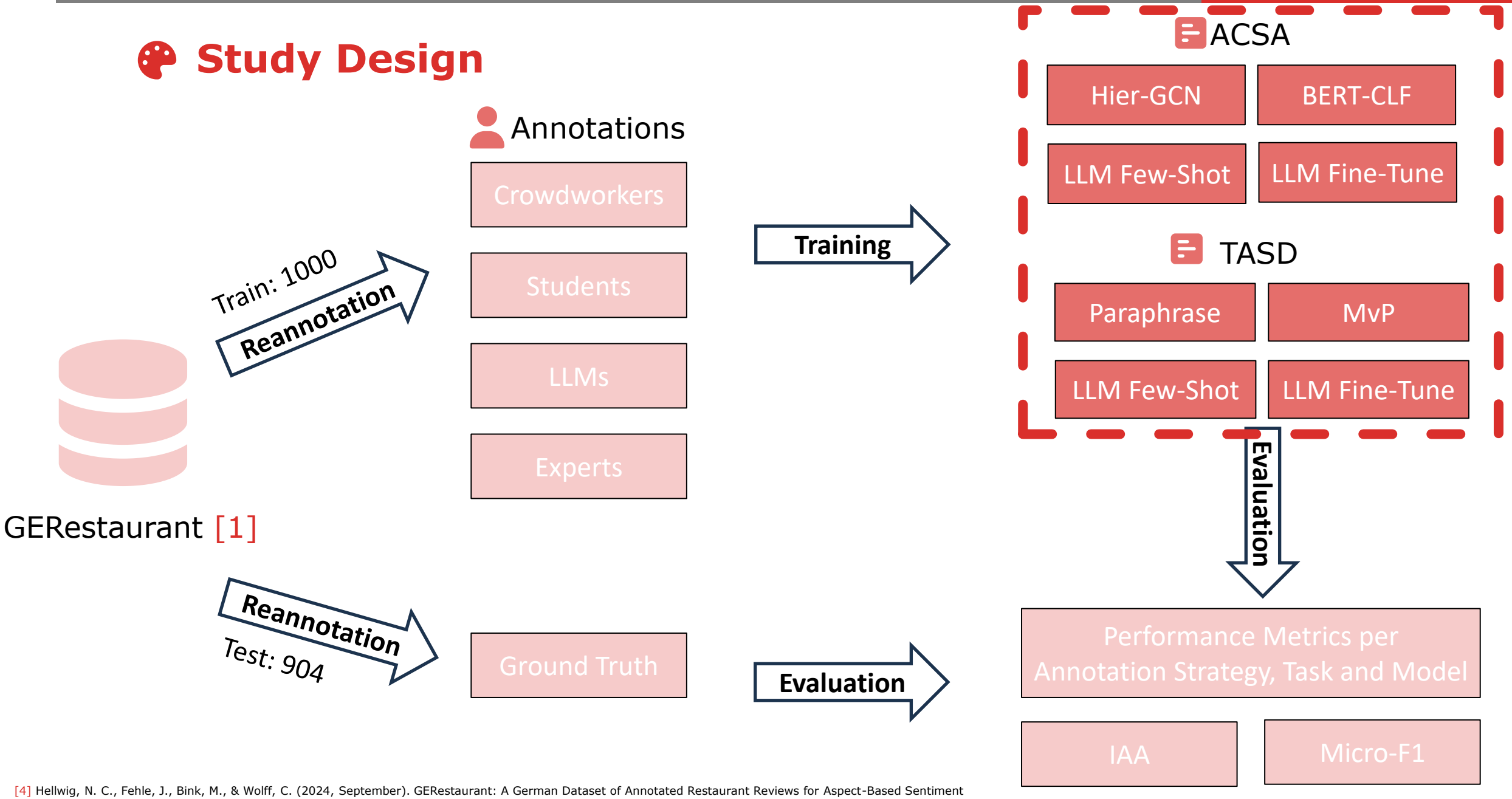
- Computer science–related students recruited via university network
- Provided with guidelines + instructional video (Label Studio workflow)
- Each batch (200 texts) annotated independently by 3 students
- Final labels derived via majority voting
- Pre-study questionnaire collected consent, background, and annotation experience (shared with crowdworkers)

Annotation: Experts

- Based on original labels from [4], adapted to revised schema
- Annotator: PhD student with ABSA expertise and involvement in original dataset curation
- Review process: accept, revise, remove
- Outcome: majority accepted, few revised, very few removed

[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

Study Design



[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

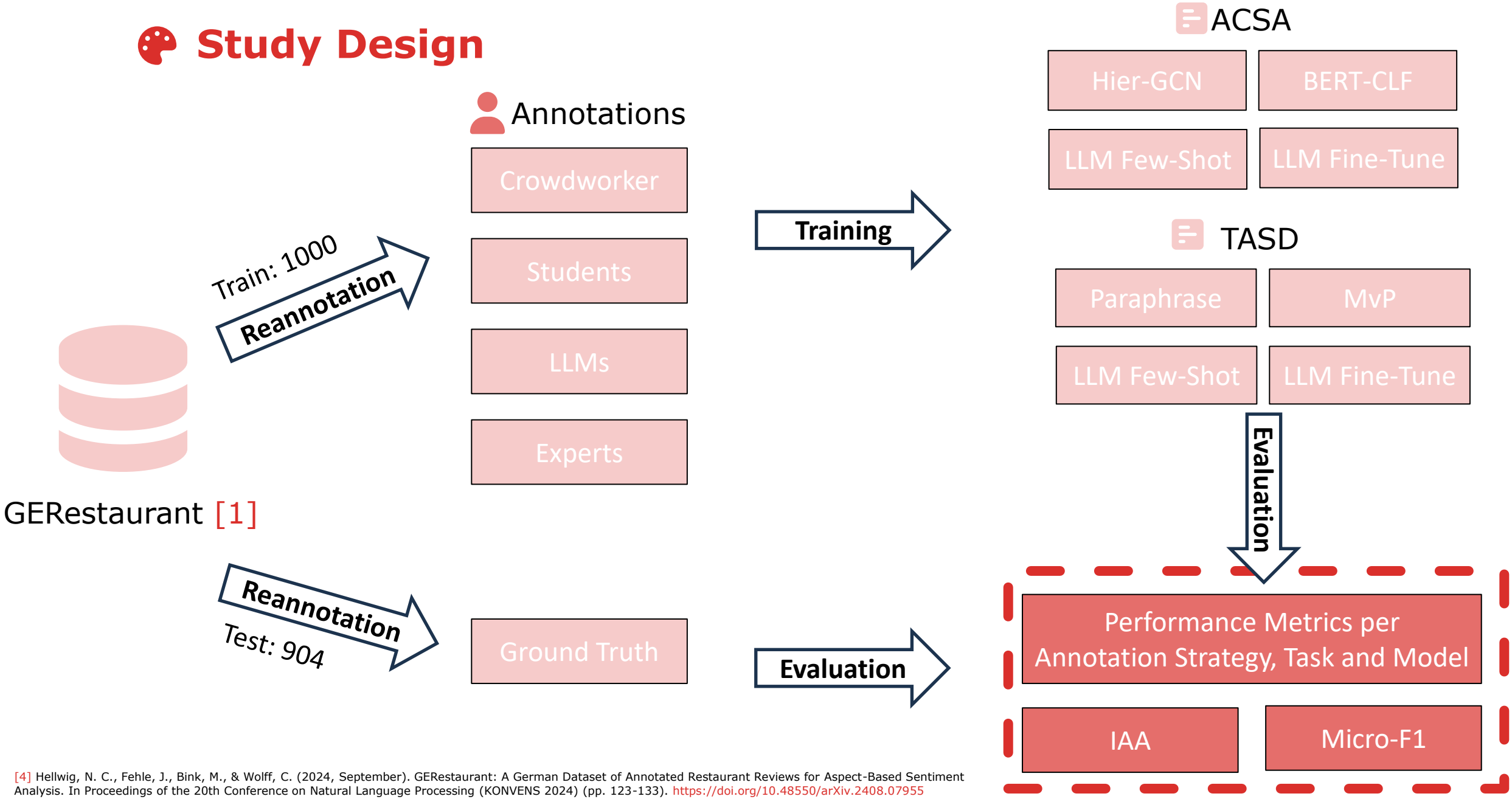
Baseline Methods

Encoder-based (classification)	Hierarchical Graph Convolutional Network [6] BERT-CLF [7]
Seq-to-Seq (text generation)	Paraphrase [8] Multi-view Prompting [9]
LLMs	Few-Shot Prompting (Gemma) [10] Instruction-based Fine-Tuning(LLaMA) [11]

[6] Cai, H., Tu, Y., Zhou, X., Yu, J., & Xia, R. (2020, December). Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th international conference on computational linguistics* (pp. 833-843). https://aclanthology.org/2020.coling-main.72.pdf
[7] Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023, September). *Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews*. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)* (pp. 202–218). https://aclanthology.org/2023.konvens-main.21.pdf
[8] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). *Aspect sentiment quad prediction as paraphrase generation*. *arXiv preprint* arXiv:2110.00796. https://doi.org/10.48550/arXiv.2110.00796
[9] Gou, Z., Guo, Q., & Yang, Y. (2023, July). *MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4380–4397). https://doi.org/10.18653/v1/2023.acl-long.240
[10] Hellwig, N., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, August). Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)* (pp. 153-172). https://aclanthology.org/2025.xllm-1.15.pdf
[11] Šmíd, J., Přibáň, P., & Kral, P. (2024, August). LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis* (pp. 63-70). https://aclanthology.org/2024.wassa-1.6.pdf

Results

Study Design



[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955

Inter-Annotator Agreement (IAA)

ACSA

- Human annotations: similar agreement, high variability
- LLMs: high agreement ≠ high quality

TASD

- Lower agreement due to higher complexity
- Crowdworkers: more errors, higher variability

Ground Truth

- Expert calibration improves consistency

	ACSA				TASD			
	GT	Crowd	Students	LLMs	GT	Crowd	Students	LLMs
Batch 1	83.93	$66.75_{\pm 11.31}$	$85.11_{\pm 1.43}$	$98.12_{\pm 0.55}$	63.33	$44.47_{\pm 16.20}$	$41.29_{\pm 17.67}$	$90.20_{\pm 2.11}$
Batch 2	88.38	$84.57_{\pm 1.43}$	$81.55_{\pm 1.01}$	$96.46_{\pm 1.10}$	70.25	$61.55_{\pm 6.41}$	$63.81_{\pm 2.34}$	$87.66_{\pm 2.82}$
Batch 3	89.66	$78.94_{\pm 2.19}$	$50.65_{\pm 25.84}$	$96.23_{\pm 1.49}$	75.78	$28.78_{\pm 20.15}$	$45.85_{\pm 7.19}$	$90.74_{\pm 2.21}$
Batch 4	88.25	$84.24_{\pm 1.60}$	$52.03_{\pm 27.54}$	$97.32_{\pm 1.27}$	74.41	$19.91_{\pm 21.97}$	$45.71_{\pm 12.52}$	$89.30_{\pm 1.95}$
Batch 5	85.78	$83.54_{\pm 2.64}$	$81.56_{\pm 1.51}$	$97.86_{\pm 0.69}$	76.95	$26.33_{\pm 26.49}$	$55.26_{\pm 9.64}$	$92.95_{\pm 1.19}$
Overall	**87.22**	**$78.95_{\pm 2.27}$**	**$63.38_{\pm 16.75}$**	**$97.20_{\pm 0.88}$**	**72.18**	**$32.38_{\pm 10.18}$**	**$50.50_{\pm 5.84}$**	**$90.22_{\pm 1.82}$**

Batch-wise IAA is measured using micro-F1 for ACSA and TASD across annotation groups. No IAA for experts (single annotator).

Inter-Annotator Agreement (IAA)

ACSA

- Human annotations: similar agreement, high variability
- LLMs: high agreement ≠ high quality

TASD

- Lower agreement due to higher complexity
- Crowdworkers: more errors, higher variability

Ground Truth

- Expert calibration improves consistency

	ACSA				TASD			
	GT	Crowd	Students	LLMs	GT	Crowd	Students	LLMs
Batch 1	83.93	$66.75_{\pm 11.31}$	$85.11_{\pm 1.43}$	$98.12_{\pm 0.55}$	63.33	$44.47_{\pm 16.20}$	$41.29_{\pm 17.67}$	$90.20_{\pm 2.11}$
Batch 2	88.38	$84.57_{\pm 1.43}$	$81.55_{\pm 1.01}$	$96.46_{\pm 1.10}$	70.25	$61.55_{\pm 6.41}$	$63.81_{\pm 2.34}$	$87.66_{\pm 2.82}$
Batch 3	89.66	$78.94_{\pm 2.19}$	$50.65_{\pm 25.84}$	$96.23_{\pm 1.49}$	75.78	$28.78_{\pm 20.15}$	$45.85_{\pm 7.19}$	$90.74_{\pm 2.21}$
Batch 4	88.25	$84.24_{\pm 1.60}$	$52.03_{\pm 27.54}$	$97.32_{\pm 1.27}$	74.41	$19.91_{\pm 21.97}$	$45.71_{\pm 12.52}$	$89.30_{\pm 1.95}$
Batch 5	85.78	$83.54_{\pm 2.64}$	$81.56_{\pm 1.51}$	$97.86_{\pm 0.69}$	76.95	$26.33_{\pm 26.49}$	$55.26_{\pm 9.64}$	$92.95_{\pm 1.19}$
Overall	**87.22**	**78.95**$_{\pm 2.27}$	**63.38**$_{\pm 16.75}$	**97.20**$_{\pm 0.88}$	**72.18**	**32.38**$_{\pm 10.18}$	**50.50**$_{\pm 5.84}$	**90.22**$_{\pm 1.82}$

Batch-wise IAA is measured using micro-F1 for ACSA and TASD across annotation groups. No IAA for experts (single annotator).

Inter-Annotator Agreement (IAA)

ACSA

- Human annotations: similar agreement, high variability
- LLMs: high agreement ≠ high quality

TASD

- Lower agreement due to higher complexity
- Crowdworkers: more errors, higher variability

Ground Truth

- Expert calibration improves consistency

	ACSA				TASD			
	GT	Crowd	Students	LLMs	GT	Crowd	Students	LLMs
Batch 1	83.93	$66.75_{\pm 11.31}$	$85.11_{\pm 1.43}$	$98.12_{\pm 0.55}$	63.33	$44.47_{\pm 16.20}$	$41.29_{\pm 17.67}$	$90.20_{\pm 2.11}$
Batch 2	88.38	$84.57_{\pm 1.43}$	$81.55_{\pm 1.01}$	$96.46_{\pm 1.10}$	70.25	$61.55_{\pm 6.41}$	$63.81_{\pm 2.34}$	$87.66_{\pm 2.82}$
Batch 3	89.66	$78.94_{\pm 2.19}$	$50.65_{\pm 25.84}$	$96.23_{\pm 1.49}$	75.78	$28.78_{\pm 20.15}$	$45.85_{\pm 7.19}$	$90.74_{\pm 2.21}$
Batch 4	88.25	$84.24_{\pm 1.60}$	$52.03_{\pm 27.54}$	$97.32_{\pm 1.27}$	74.41	$19.91_{\pm 21.97}$	$45.71_{\pm 12.52}$	$89.30_{\pm 1.95}$
Batch 5	85.78	$83.54_{\pm 2.64}$	$81.56_{\pm 1.51}$	$97.86_{\pm 0.69}$	76.95	$26.33_{\pm 26.49}$	$55.26_{\pm 9.64}$	$92.95_{\pm 1.19}$
Overall	**87.22**	**78.95**$_{\pm 2.27}$	**63.38**$_{\pm 16.75}$	**97.20**$_{\pm 0.88}$	**72.18**	**32.38**$_{\pm 10.18}$	**50.50**$_{\pm 5.84}$	**90.22**$_{\pm 1.82}$

Batch-wise IAA is measured using micro-F1 for ACSA and TASD across annotation groups. No IAA for experts (single annotator).

Results: ACSA

- Expert data performs best overall, but differences are small across datasets
- LLM-based models outperform classical baselines; best results with expert + LLMs
- Few-shot setting: student data performs best (competitive non-expert annotations)

Method	Crowd	Students	LLMs	Experts
BERT-CLF	76.99	77.81	77.44	**78.26**
Hier-GCN	79.66	78.97	79.13	**79.78**
Gemma FS	86.03	**86.43**	85.60	86.29
LLaMA FT	85.64	85.71	84.85	**86.39**

(a) Aspect Category Sentiment Analysis (ACSA)

Micro-F1 (avg. 5 seeds) for ACSA across annotation sources.

Results: ACSA

- Expert data performs best overall, but differences are small across datasets
- LLM-based models outperform classical baselines; best results with expert + LLMs
- Few-shot setting: student data performs best (competitive non-expert annotations)

Method	Crowd	Students	LLMs	Experts
BERT-CLF	76.99	77.81	77.44	**78.26**
Hier-GCN	79.66	78.97	79.13	**79.78**
Gemma FS	86.03	**86.43**	85.60	86.29
LLaMA FT	85.64	85.71	84.85	**86.39**

(a) Aspect Category Sentiment Analysis (ACSA)

Micro-F1 (avg. 5 seeds) for ACSA across annotation sources.

Results: ACSA

- Expert data performs best overall, but differences are small across datasets
- LLM-based models outperform classical baselines; best results with expert + LLMs
- Few-shot setting: student data performs best (competitive non-expert annotations)

Method	Crowd	Students	LLMs	Experts
BERT-CLF	76.99	77.81	77.44	**78.26**
Hier-GCN	79.66	78.97	79.13	**79.78**
Gemma FS	86.03	**86.43**	85.60	86.29
LLaMA FT	85.64	85.71	84.85	**86.39**

(a) Aspect Category Sentiment Analysis (ACSA)

Micro-F1 (avg. 5 seeds) for ACSA across annotation sources.

Results: TASD

- Expert annotations achieve best overall performance; fine-tuned LLM is strongest
- Crowdworkers dataset performs worst; student and LLM datasets are comparable
- TASD shows larger performance gaps than ACSA due to higher complexity

Method	Crowd	Students	LLMs	Experts
Paraphrase	52.77	57.33	57.37	**61.65**
MvP	51.29	56.83	60.65	**64.01**
Gemma FS	58.56	62.28	**65.58**	63.38
LLaMA FT	65.46	69.33	66.24	**71.47**

(b) Target Aspect Sentiment Detection (TASD)

Micro-F1 (avg. 5 seeds) for TASD across annotation sources.

Results: TASD

- Expert annotations achieve best overall performance; fine-tuned LLM is strongest
- Crowdworker dataset performs worst; student and LLM datasets are comparable
- TASD shows larger performance gaps than ACSA due to higher complexity

Method	Crowd	Students	LLMs	Experts
Paraphrase	52.77	57.33	57.37	**61.65**
MvP	51.29	56.83	60.65	**64.01**
Gemma FS	58.56	62.28	**65.58**	63.38
LLaMA FT	65.46	69.33	66.24	**71.47**

(b) Target Aspect Sentiment Detection (TASD)

Micro-F1 (avg. 5 seeds) for TASD across annotation sources.

Conclusion

Limitation & Future Work

Limitations

- Crowdworkers annotations: scalable but costly
- Student annotation: time-intensive process
- LLM annotations: bias + high compute needs
- Experts: difficult to recruit & time-intensive
- Human annotation: potential unseen use of external tools

Future Work

- Extend to other domains
- Combine LLM + expert annotations

Summary

Background

- Limited availability of German ABSA datasets
- Lack of studies on complex annotation quality such as ABSA

Research Goals

- Investigate the impact of annotation type on ABSA dataset quality

Annotation Sources

- Crowdsourcing
- Large Language Models
- Students
- Experts

Key Result

- LLMs are a scalable alternative, but expert quality still sets the upper bound
- Clear guidelines matter for annotation quality

Contact

GitHub

https://github.com/NiklasDonhauser/absa-annotation-quality

Mail

Niklas.Donhauser@ur.de

Sources

[1] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature. https://hyse.org/pdf/SentimentAnalysis-and-OpinionMining.pdf
[2] Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review, 49, 100576. https://doi.org/10.1016/j.cosrev.2023.100576
[3] Singhi, V., Chauhan, C., & Soni, P. K. (2024, April). Exploring Progress in Aspect-based Sentiment Analysis: An In-depth Survey. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-10). IEEE. https://doi.org/10.1109/I2CT61223.2024.10543612
[4] Hellwig, N. C., Fehle, J., Bink, M., & Wolff, C. (2024, September). GERestaurant: A German Dataset of Annotated Restaurant Reviews for Aspect-Based Sentiment Analysis. In Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024) (pp. 123-133). https://doi.org/10.48550/arXiv.2408.07955
[5] Hellwig, N. C., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, May). *Do we still need human annotators? Prompting large language models for aspect sentiment quad prediction*. arXiv. https://doi.org/10.48550/arXiv.2502.13044
[6] Cai, H., Tu, Y., Zhou, X., Yu, J., & Xia, R. (2020, December). Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th international conference on computational linguistics* (pp. 833-843). https://aclanthology.org/2020.coling-main.72.pdf
[7] Fehle, J., Münster, L., Schmidt, T., & Wolff, C. (2023, September). *Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews*. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)* (pp. 202–218). https://aclanthology.org/2023.konvens-main.21.pdf
[8] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). *Aspect sentiment quad prediction as paraphrase generation*. *arXiv preprint* arXiv:2110.00796. https://doi.org/10.48550/arXiv.2110.00796
[9] Gou, Z., Guo, Q., & Yang, Y. (2023, July). *MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4380–4397). https://doi.org/10.18653/v1/2023.acl-long.240
[10] Hellwig, N., Fehle, J., Kruschwitz, U., & Wolff, C. (2025, August). Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)* (pp. 153-172). https://aclanthology.org/2025.xllm-1.15.pdf
[11] Šmíd, J., Přibáň, P., & Kral, P. (2024, August). LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis* (pp. 63-70). https://aclanthology.org/2024.wassa-1.6.pdf

Appendix

Dataset Distribution

- Annotation sources: crowdworkers, students, LLMs, experts and test set
- Test set slightly smaller