SciBERT Meets Contrastive Learning: A Solution for Scientific Hallucination Detection

Carla Crivoi[1] **Ana-Sabina Uban**[1,2]
[1]Faculty of Mathematics and Computer Science, University of Bucharest
[2]Human Language Technologies Research Center, University of Bucharest
crivoicarla02@gmail.com, auban@fmi.unibuc.ro

Abstract

Large language models are increasingly used to synthesize scientific literature, yet they remain prone to *hallucination* — claims that are linguistically fluent but lack support in the cited sources. We tackle hallucination detection in the SciHal 2025 challenge by augmenting SciBERT with Triplet and InfoNCE contrastive objectives in addition to cross-entropy classification. The system achieves validation macro-F_1 scores of 0.626 ± 0.004 on the coarse-grained hallucination detection task (Sub-task 1) and 0.632 ± 0.012 on the fine-grained detection task (Sub-task 2), exceeding a plain SciBERT baseline by more than three points. The official blind test set scores reach macro-F_1 scores of 0.51 and 0.43 for Sub-tasks 1 and 2, respectively, securing fifth place in both leaderboards. Confusion matrix analysis shows that contrastive learning markedly improves majority classes, whereas sparse categories, especially *Missing Information*, remain challenging despite aggressive attempts to mitigate class imbalance.

1 Introduction

Large language models (LLMs) such as ChatGPT (OpenAI, 2023) are increasingly used to support academic research by answering domain-specific questions and summarising scientific content. While their outputs are often fluent and persuasive, they may introduce statements that are not grounded in the source material — a phenomenon known as *hallucination*. Detecting hallucinated claims is especially difficult in scientific domains, where language is highly specialised and reference documents are lengthy.

In this work we present a contrastive-learning solution based on SciBERT (Beltagy et al., 2019) for hallucination detection in scientific answers. Our contributions are three-fold: (i) a systematic analysis of the SciHal corpus that highlights the linguistic and structural challenges of the task, (ii) a multi-objective optimisation scheme that couples classification with two contrastive losses, and (iii) discussion and analysis of results and errors, including confusion matrix diagnostics, demonstrating the effectiveness of the proposed model.

2 Related Work

Early studies of factual consistency focused on abstractive summarisation, where hallucinations degrade summary quality. Maynez et al. (2020) showed that even state-of-the-art models hallucinate frequently, motivating automatic detection methods such as Question Answering (QA)-based factuality probes (Kryściński et al., 2020). With the advent of large language models (LLMs) like GPT-3, hallucinations have been documented in open-domain Question Answering (Ji et al., 2023) and conversational agents (Thoppilan et al., 2022). Most approaches frame hallucination detection as either an entailment problem, requiring reference retrieval and contradiction detection, or a generation-probability anomaly task.

Contrastive objectives have proven effective at learning semantically meaningful representations from limited supervision (Chen et al., 2020). In factuality research, Liu et al. (2022) applied supervised contrastive loss to claim verification, achieving gains over cross-entropy-only training. Yuan et al. (2022) employed Information Noise-Contrastive Estimation (InfoNCE) to align biomedical entity mentions with definitions, improving downstream question answering performance. For hallucination mitigation, Shi et al. (2023) used retrieval-augmented contrastive tuning to discourage unsupported generations, while Deng et al. (2024) introduced dual-encoder contrastive pre-training to rank evidence passages. Our work differs by combining *two* contrastive losses: Triplet and InfoNCE with a cross-entropy objective inside a SciBERT backbone, targeting both coarse and fine-grained hallucination labels in scientific texts.

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 336–343

3 Task Description

The Hallucination Detection for Scientific Content (SciHal) shared task addresses a challenge in the use of generative AI-powered academic research assistants: the detection of hallucinated claims in automatically generated scientific answers. These hallucinations—claims unsupported by reliable sources—undermine the trustworthiness of AI-generated scientific content.

The task is formulated as a multi-label classification problem, where participants are required to assess the factual consistency of claims generated in response to research-related questions. For each instance, participants are provided with: a question related to scientific research, a summarized answer produced by a generative AI system, an extracted claim from that answer, and the corresponding reference abstracts cited in support of the summary.

Participants must determine whether each claim is factually supported or hallucinatory based on the provided reference materials. The task is divided into two sub-tasks: coarse-grained hallucination detection, and fine-grained hallucination detection.

3.1 Sub-task 1: Coarse-grained Hallucination Detection

In the first sub-task, each claim must be classified into one of the following categories:

- **Entailment**: the claim is supported by the references.
- **Unverifiable**: the claim cannot be verified using the provided references.
- **Contradiction**: the claim contradicts information in the references.

3.2 Sub-task 2: Fine-grained Hallucination Detection

The second sub-task requires a more fine-grained analysis of hallucination types. Each claim must be categorized as one of the following: Entailment, Unrelated and unverifiable, Related but unverifiable, Misrepresentation, Missing information, Numeric error, Entity error, Opposite meaning.

3.3 Evaluation Metrics

We evaluate models with the macro F_1 score, which assigns equal weight to every class by averaging their per-class F_1 values, irrespective of class frequency. To provide a more granular picture of errors, we also include confusion matrices for each sub-task, detailing how predictions are distributed across the true labels.

4 Methodology

4.1 Dataset and Split Strategy

The official SciHal release provides 3,592 labelled instances for Sub-task1 and 4,092 for Sub-task2. Following the shared-task protocol we adopt an 85:15 split, corresponding to 3,053 / 539 (train / validation) examples for Sub-task1 and 3,478 / 614 examples for Sub-task2. The test sets were not released to the participants, but the submissions were evaluated on 50% on the test data using the same metrics to obtain the team rankings for both sub-tasks.

4.2 Data Analysis

Tables 1-3 summarise descriptive statistics for the dataset. These numbers highlight linguistic and structural challenges: input sequences vary substantially in length and claims are concise, whereas references are much longer. A lexical overlap analysis provides further evidence: the average Jaccard coefficient (da F. Costa, 2021) between lemmatised claim and reference token sets is 0.092 (minimum 0.000; maximum 0.474), confirming that surface-form overlap is generally low.

Field	Max C	Min C	Avg C
Question	269	15	80.06
Claim	705	28	256.18
Answer	5649	897	3426.23
Reference	19375	190	2046.49

Table 1: Character count statistics across text fields.

Field	Avg W	Max W	Min W
Question	11.24	37	2
Claim	36.02	104	4
Answer	465.18	757	133
Reference	299.91	2824	30

Table 2: Word count statistics across text fields.

4.3 Proposed Solution

Our system tackles hallucination detection by fine-tuning SciBERT within a contrastive-learning paradigm. The network features a dual-head design: a classification branch with two dense lay-

Field	Avg S
Question	1.01
Claim	1.70
Answer	22.00
Reference	15.66

Table 3: Average sentence count per field.

ers with layer normalization, dropout, and a softmax output, and a projection branch consisting of a two-layer MLP with ReLU and dropout whose L_2-normalised embeddings serve the contrastive objectives. We pool the final hidden states by concatenating the `[CLS]` vector with the mean of all token embeddings, yielding a hybrid representation that feeds both heads.

Optimisation relies on a composite loss,

$$\mathcal{L} = 0.3(\mathcal{L}_{Triplet} + \mathcal{L}_{InfoNCE}) + 0.7\,\mathcal{L}_{CE},$$

where Triplet Loss (Schroff et al., 2015) enforces distance constraints between positive and negative claim-reference pairs, *InfoNCE Loss* (Oord et al., 2018) promotes high cosine similarity among positives, and *Cross-Entropy Loss* (Bishop, 2006) supplies the multi-class signal. A grid search confirmed that the 30:70 contrastive classification weighting gives the best validation performance.

During training we adopt differential learning rates: 5×10^{-5} for the encoder, and 5×10^{-4} for the task-specific layers cosine annealing with a 15% warm-up, early stopping (maximum 25 epochs), and gradient clipping at an L_2-norm of 5 to prevent exploding updates.

We employ a weighted loss in order to mitigate class imbalance. Class weights are computed as

$$w_i = \frac{N}{K \cdot n_i}, \tag{1}$$

where N is the total number of samples, K the number of classes, and n_i the frequency of class i.

5 Experiments

5.1 Sub-task 1

5.1.1 Dataset and Preprocessing

The SciHal dataset comprises 3,592 labeled instances with class distribution: *contra* (1,369), *entail* (1,333), and *unver* (890). We employed an 85:15 train-validation split, yielding 3,053 training and 539 validation examples. Class weights were computed to mitigate the observed label imbalance during training.

5.1.2 Results

The model reached its peak validation performance at epoch 5 with a macro F_1 of 0.626 ($\pm$0.004 across five runs); the corresponding per-class F_1 scores were 0.673 for *contra*, 0.600 for *entail*, and 0.591 for *unver*.

In this run, 1,000 validation instances were classified as follows: *entail* (521), *contra* (251), and *unver* (228). The mean prediction confidence, computed as the probability output by the model for the predicted class, was 0.892, with only 17 predictions falling below a 0.60 threshold; class-specific average confidences were 0.923 for *unver*, 0.888 for *contra*, and 0.881 for *entail*.

Figure 1 presents the normalised confusion matrix obtained from a *fresh* evaluation run using the same experimental settings. Small numerical deviations from the previous report reflect the non-deterministic nature of stochastic optimisation and mini-batch sampling.

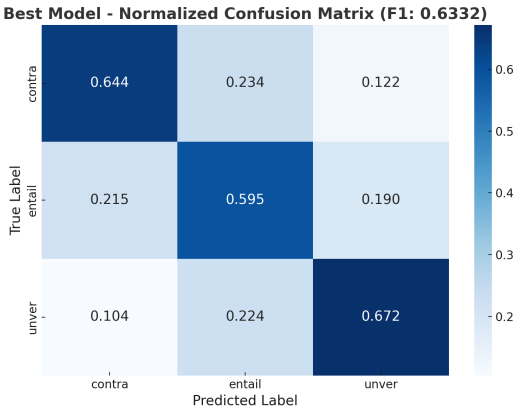

Figure 1: Normalised confusion matrix for the best validation checkpoint for Sub-task 1.

The confusion matrix shows that the *average misclassification rate*, defined as the sum of all off-diagonal cell counts divided by the total number of validation instances, is 0.1815. The *standard deviation* of these off-diagonal error proportions is 0.0505, indicating a moderate spread: while roughly 18% of inputs are assigned to an incorrect class, the class-to-class variability rarely exceeds $\pm$5 percentage points.

On the official blind test set released by the SciHal 2025 organisers our final submission, trained with the configuration described above, attained a macro-F_1 score of 0.51, which placed us fifth out of all participating teams.

5.2 Sub-task 2

5.2.1 Dataset and Preprocessing

Sub-task 2 employs the extended SciHal corpus of 3,592 annotated instances covering eight hallucination categories. The data were randomly partitioned in an 85:15 ratio, resulting in 3,053 training examples and 539 validation examples. Because the class distribution is heavily skewed (categories such as *missinfo*, *numerr*, and *unrelunvef* are markedly under-represented), we adopt inverse-frequency weighting on the training split only. The resulting weights are shown in Table 4.

Class label	Train examples	Weight w_i
Entail	1333	0.337
Related-Unverifiable	738	0.608
Opposite Meaning (negat)	625	0.718
Misrepresentation	395	1.137
Entity Error	174	2.580
Unrelated-Unverifiable	152	2.954
Numeric Error	116	3.871
Missing Information	59	7.610

Table 4: Class frequencies in the training split (3,478 instances) and the inverse-frequency weights used during optimisation for Sub-task 2.

5.2.2 Training Configuration and Results

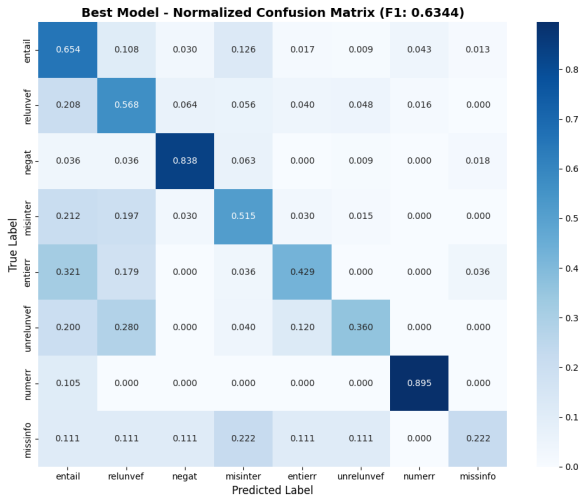

Figure 2: Normalised confusion matrix for the best validation checkpoint for Sub-task 2.

The experimental setup mirrors that of Sub-task 1; the only architectural difference is the output softmax now spans eight classes instead of three. Validation performance rose steadily and peaked at a macro-F_1 of 0.632 ($\pm$0.012 across five runs) on epoch 19.

The confusion matrix in Figure 2 further shows that the *missinfo* category, despite receiving the largest class weight, remains; its under-representation renders it the most difficult label to learn, illustrating that even aggressive re-weighting cannot fully offset data sparsity.

For Sub-task 2 the same model configuration achieved a macro-F_1 of 0.43 on the shared-task test set, securing fifth place in the final leaderboard. While performance naturally drops in the more fine-grained, eight-class scenario, the result demonstrates that our contrastive SciBERT approach remains competitive even when the label space is enlarged and class imbalance becomes more pronounced.

Additional experiments and results are reported in the Appendix, including results with a vanilla fine-tuned SciBERT model using only the cross-entropy objective, which obtains poorer validation results were poorer than our final approach.

6 Conclusion

We have introduced a contrastive-learning extension of SciBERT for detecting hallucinated claims in AI-generated scientific answers. Jointly optimising cross-entropy with Triplet and InfoNCE losses yields consistent gains on both coarse- and fine-grained settings of the SciHal 2025 benchmark, outperforming an unweighted baseline and a purely cross-entropy model. The improvement is most pronounced for majority and medium-frequency labels, confirming that semantic alignment objectives complement token-level supervision. Nonetheless, the model still struggles with the under-represented classes, indicating that re-weighting alone cannot fully offset data scarcity.

Future work could improve performance by model updates along three possible axes. First, coupling the encoder with a retrieval-compression module that distills each reference into a handful of salient sentences could help by thereby shortening inputs while preserving key evidence. Second, we intend to introduce a curriculum that over-samples rare labels and structurally complex claims early in training, then relaxes the sampling schedule as the model stabilizes. Third, we will examine whether parameter-efficient fine-tuning of substantially larger transformer backbones improves robustness, especially on the sparsest categories, without incurring prohibitive computational cost.

Limitations

Our approach has several practical and methodological limitations. First, all experiments were conducted using a single NVIDIA Tesla P100 GPU, which constrained the batch size and training speed, especially during contrastive learning. Due to memory limitations, we relied on models from the BERT family, which support a maximum input length of 512 tokens. This likely prevented the model from accessing the full context in cases where the reference abstracts were lengthy or complex.

Another key limitation is the relatively small size of the training dataset. While sufficient for fine-tuning, the number of examples is limited from the perspective of large language models (LLMs), increasing the risk of overfitting and limiting generalization. This was especially evident for underrepresented labels in Sub-task 2, where performance gains plateaued early. More data and better-balanced class distributions would likely improve robustness.

Lastly, our model processes claims and references independently at the input level, without explicitly modeling document structure or reasoning chains. Incorporating more advanced context handling or retrieval-augmented methods could help mitigate this in future work.

Ethics Statement

This work focuses on improving the factual reliability of AI-generated scientific content by detecting hallucinated claims. Our intention is to support responsible use of large language models (LLMs) in academic research, not to automate or replace scientific reasoning. We recognize that LLMs may still introduce errors or biased outputs, and systems built on top of them should always be used with human oversight.

We used publicly released data provided by the SciHal 2025 shared task organizers, and did not collect or annotate any additional human data. No personally identifiable information (PII) was involved. Our models were trained and evaluated only for research purposes, and we do not deploy them in production systems.

We also acknowledge the computational costs of training large models. While we used relatively modest hardware (a single P100 GPU), future work should continue to consider the environmental impact of large-scale training.

Finally, we emphasize that hallucination detection is not a solved problem, and there is a risk that users may overtrust partially automated systems. Clear communication of model limitations and transparency in design choices are essential to ensure ethical deployment.

Acknowledgements

This research was partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906, and by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *EMNLP*.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR.

Luciano da F. Costa. 2021. Further generalizations of the jaccard index.

Xiang Deng, Han Zhang, Wenhao Yu, Clare Lee, and Mohit Bansal. 2024. Factscore 2.0: Dual-encoder contrastive pre-training for factual consistency. *Transactions of the Association for Computational Linguistics*, 12:1–19.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Zhenhao Liu, Haoran Xu, and Huan Sun. 2022. Fine-grained fact verification with supervised contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1235–1247. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919, Online. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. Gpt-4 technical report. ArXiv:2303.08774.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Weizhe Shi, Shijie Wu, Xinyun Chen, and Xiang Ren. 2023. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1600–1618. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Xuanji Yuan, Tao Shen, Shawn Tan, and Min-Yen Kan. 2022. Improving biomedical question answering via contrastive learning in biobert. In *Proceedings of the 21st Workshop on Biomedical Language Processing (BioNLP)*, pages 156–167. Association for Computational Linguistics.

A Additional Training Statistics

To assess the stability of our optimisation procedure, we repeated each model training five times with different random seeds and report the best-checkpoint macro F_1.

A.1 Sub-task 1

Table 5 summarises statistics for the coarse-grained results.

Statistic	Macro F_1
Mean	0.628
Standard deviation	0.004

Table 5: Validation macro F_1 across five independent training runs for Sub-task 1.

A.2 Sub-task 2

Table 6 reports statistics for the fine-grained, eight-class setting.

Statistic	Macro F_1
Mean	0.632
Standard deviation	0.012

Table 6: Validation macro F_1 across five independent training runs for Sub-task 2.

A.3 Training Dynamics Sub-task 1

Figure 3 illustrates macro F_1 evolution across epochs, while Figure 4 shows per-class F_1 trajectories.

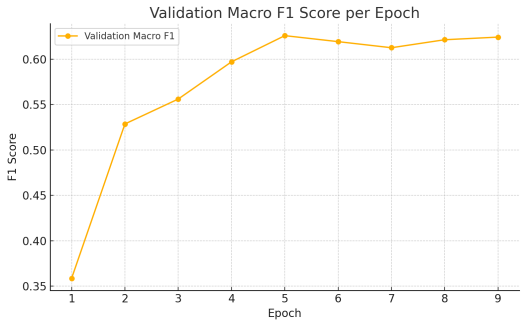

Figure 3: Macro F_1 Score Evolution

A.4 Training Dynamics Sub-task 2

Figure 5 presents macro F1 evolution across training epochs, while Figure 6 illustrates per-class F1 trajectories for representative categories.

B SciBERT Baseline (No Contrastive Learning or Class Weights)

To establish an absolute reference point, we fine-tuned a vanilla SciBERT model using only the cross-entropy objective and *no* class weighting or

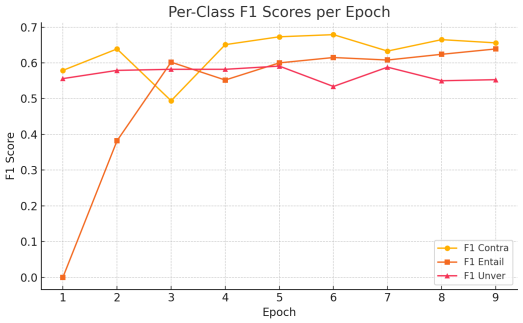

Figure 4: Per-Class F_1 Score Evolution

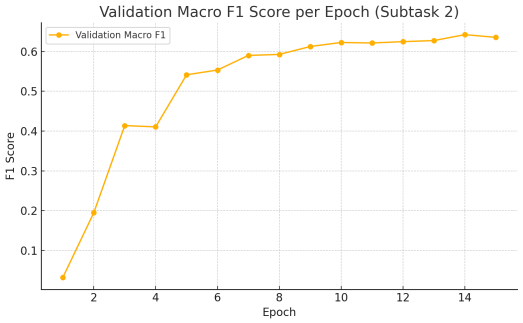

Figure 5: Validation Macro F1 Score Evolution (Subtask 2)

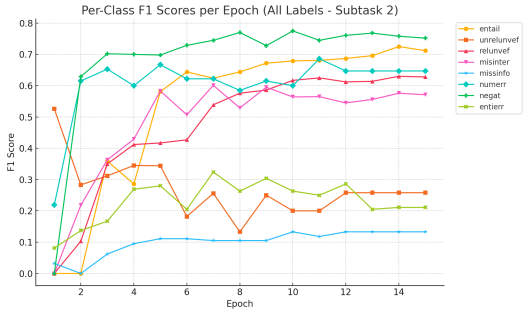

Figure 6: Per-Class F1 Score (Sub-task 2)

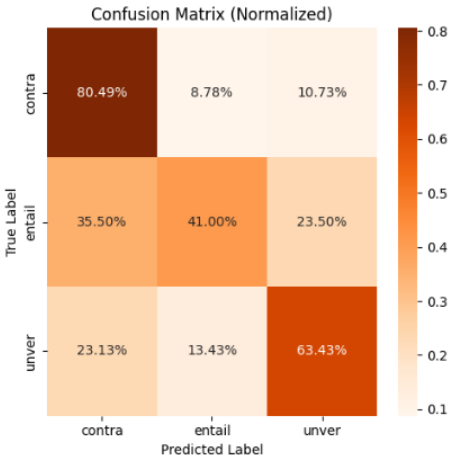

Figure 7: Confusion matrix for the SciBERT baseline on Sub-task 1 (validation macro-$F_1 = 0.601$).

contrastive losses. Training was performed with early stopping (maximum 10 epochs) and a learning rate of 2×10^{-5}. Table 7 reports the resulting macro-F_1 scores, while Figures 7 and 8 show the corresponding confusion matrices. Overall, our final approach using class weighting and contrastive loss seems to obtain improvements compared to the baseline for most classes, while the most notable difference is in the rare classes, such as *Missing Information*, for which the simple baseline does not manage to classify almost any examples correctly.

	Macro F_1	
	Sub-task 1	Sub-task 2
SciBERT (baseline)	0.601	0.586

Table 7: Validation macro F_1 for the SciBERT baseline trained without contrastive objectives or class weighting.

B.1 Confusion Matrix for SciBERT Baseline for Sub-task 1.

Figure 7 shows the confusion matrix of the vanilla SciBERT baseline, which yields a validation macro-F_1 of 0.601 on Sub-task 1.

B.2 Confusion Matrix for SciBERT Baseline for Sub-task 2.

Figure 8 displays the confusion matrix of the SciBERT baseline, which attains a validation macro-F_1 of 0.586 on Sub-task 2.

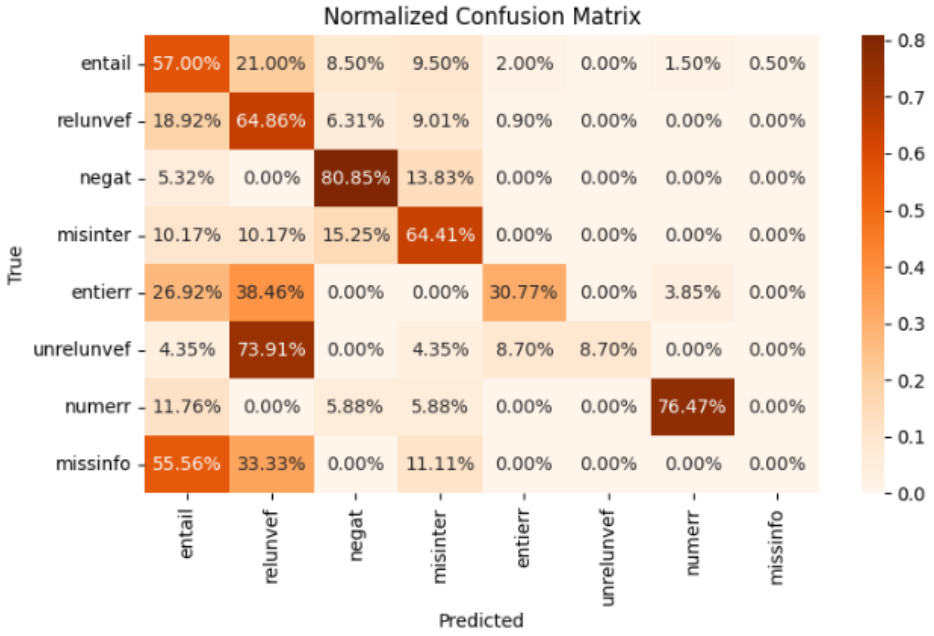

Figure 8: Confusion matrix for the SciBERT baseline on Sub-task 2 (validation macro-$F_1 = 0.586$).