# Modernizing GENIA dataset for biomedical event extraction: a preliminary reannotation experiment

Dina Stein[1,*], Darya Kuzmenko[1], Vitaly Romanov[1], Kacper Ogórek[1], Aleksander Tarelkin[1] and Denis Stepanov[1]

[1]*JetBrains s.r.o.*

**Abstract**

Established datasets for biomedical event instruction are often over a decade old and show performance limitations. This study presents a reannotation experiment to modernize the widely-used GENIA dataset. We reannotated 50 abstracts through collaboration between computational linguists and biomedical experts. The reannotated dataset was evaluated for biological plausibility by four experts and for model performance using BERT- and LLM-based architectures. Results show improvements in biological plausibility, evidenced by an increase in expert agreement with annotation for event types (from 0.72 to 0.86), accompanied by a gain in inter-rater agreement. Model performance was maintained or improved, with LLM F1-score increasing from 0.70 to 0.74. These findings demonstrate that systematic reannotation can enhance both biological validity and computational tractability, providing a foundation for modernizing biomedical event extraction datasets.

**Keywords**

biomedical event extraction, GENIA dataset, dataset modernization

## 1. Introduction

Biomedical event extraction (BioEE) is a natural language processing task aimed at extracting structured information from biomedical texts. Crucially, however, the most popular datasets used in the BioEE field were created more than 10 years ago and were rarely revisited [1]. This justifies the need for an update to incorporate new knowledge into the annotation, and to address the issues in the original datasets.

One of the main datasets in the field of BioEE is the GENIA dataset [1]. The performance of various models on this dataset for the event trigger classification task, however, seems to be capped at a moderate F1-score under 0.7 [2]. Our preliminary study suggested that issues with annotation consistency may underlie this limitation on performance.

The present study is aimed at testing whether updating the annotation could improve the annotation consistency and bring it into better correspondence with the current state of biomedical knowledge.

## 2. GENIA reannotation and evaluation

The original GENIA dataset annotation includes events, named entities, coreferences, and meta-knowledge. For the present study, we will only consider the event triggers, that is, the words in the text that indicate the presence of a biological event.

Our experiment consisted of two steps: reannotation and evaluation. The results of the reannotation were changes in the event type ontology, as well as corrections, removing, adding, or changing the event annotations to both fix issues and improve consistency. In total there were 1639 event annotations in the original 50 abstracts, and 2153 after reannotation, with 587 being added, 73 deleted, 1127 unchanged, 157

*17th International SWAT4HCLS Conference, March 23–26, 2026, Amsterdam, The Netherlands*

The authors declare the following use of generative AI: Generative AI was used as one of the prediction approaches in the present experiments and to reformulate parts of the text. All AI-generated content was reviewed and edited by the authors.

*Corresponding author.

dina.stein@jetbrains.com (D. Stein)

0009-0007-6019-3333 (D. Stein); 0000-0003-3772-0039 (V. Romanov)

changed in terms of selected text and 317 changed in terms of event type. Additionally, the event types were given definitions in accordance with the current Gene Ontology Biological Process definitions [3].

The evaluation step was aimed at checking two key aspects: (a) the biological plausibility of the annotation; and (b) the performance of two model architectures (BERT- and LLM-based) on either dataset version.

The biological plausibility was assessed by asking 4 biomedical specialists. The experts were presented with 5 event examples for each event type from both original and reannotated dataset.

The model performance was evaluated for either model architecture on the original or reannotated dataset. The BERT-based model used PubMedBERT as the backbone [4], with additional encoder layers and a final classification layer. Due to small re-annotated dataset size, the BERT-based model was pre-trained on the original GENIA dataset until convergence and then fine-tuned on re-annotated training subset. For the LLM-based approach we used Claude 4 Sonnet in a many-shot manner, the model being prompted with examples from either dataset version. Apart from examples, the prompt included a general instruction and the definition of the events.

Our evaluation results suggest that the changes we introduced during reannotation were beneficial in terms of biological plausibility. The agreement rate for event type improved (0.72 vs 0.86), as well as the inter-rater agreement for event types (0.47 vs 0.59 Prediger's Kappa [5]). At the same time, the reannotation maintained or improved the model performance for either BERT- or LLM-based architecture, as summarized in Table 1.

**Table 1**
The performance for event trigger prediction on the test set of the experimental subset on the original and reannotated dataset for BERT- and LLM-based model architectures.

| Architecture | F1 | | Precision | | Recall | |
|---|---|---|---|---|---|---|
| | original | reannotated | original | reannotated | original | reannotated |
| BERT-based | 0.7776 | 0.7758 | 0.7213 | 0.7083 | 0.8437 | 0.8578 |
| LLM-based | 0.7031 | 0.7444 | 0.5979 | 0.6370 | 0.8536 | 0.8955 |

A detailed error analysis revealed that on changed examples BERT F1 score increased by 0.14 and LLM F1 score by 0.20. We also observed that BERT-based approach could not successfully learn the added examples given the limited data: F1 for added examples was below average on reannotated dataset by 0.07; while for LLM the novel examples showed comparable-to-overall performance.

Overall, we show that harmonizing data annotation schemes might help improve biological plausibility of the annotation, while simultaneously improving or keeping the same level of model performance.

## References


[1] G. Frisoni, G. Moro, A. Carbonaro, A survey on event extraction for natural language understanding: Riding the biomedical literature wave, IEEE Access 9 (2021) 160721–160757. doi:10.1109/ACCESS.2021.3130956.

[2] H. Yuan, S. C. Hui, H. Zhang, A structure-aware generative model for biomedical event extraction, arXiv preprint arXiv:2408.06583 (2024). doi:10.48550/arXiv.2408.06583.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. doi:10.1038/75556, the Gene Ontology Consortium.

[4] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23. doi:10.1145/3458754.

[5] R. L. Brennan, D. J. Prediger, Coefficient kappa: Some uses, misuses, and alternatives, Educational and Psychological Measurement 41 (1981) 687–699. doi:10.1177/001316448104100307.