BNLP: A Text Annotation Platform for Quality Control of LLM-Generated Annotations

Xinhao Zhuang[1*], **Qiongyu Tian**[1*], **Yalin Chen**[3], **Tianle Xin**[1],
Yongyong Fu[2], **Yunchao Ling**[1†], **Guoqing Zhang**[1†]
[1]Shanghai Institute of Nutrition and Health, University of the Chinese Academy of Sciences
[2]Hangzhou Institute for Advanced Study, University of the Chinese Academy of Sciences
[3]Fudan University
{zhuangxinhao2019,tianqiongyu2024,gqzhang}@sinh.ac.cn

Abstract

High-quality annotated data is crucial for NLP, yet manual annotation is costly and difficult to scale in low-resource settings. Large Language Models (LLMs) have demonstrated strong zero-shot and few-shot generalization in NLP tasks, but existing annotation tools either lack LLM support or use LLMs only as one-off pre-annotation engines, without incorporating collaboration or quality control, compromising data reliability. We present BNLP, a text annotation platform that embeds LLM-assisted labeling into a quality-aware collaborative workflow. BNLP treats LLM outputs as intermediate, revisable states and integrates multi-role collaboration, iterative review cycles, and consistency analysis to enable continuous quality monitoring while preserving efficiency gains. BNLP also natively supports AI-ready formats such as Excel and JSON, ensuring seamless data flow from manual annotation to model training. Experiments show that BNLP reduces annotation time by 74.3% and improves annotation quality by 11.6% over purely manual annotation in LLM-assisted settings.[1]

1 Introduction

High-quality annotated data remains fundamental to building robust AI models. With the rapid rise of generative Large Language Models (LLMs), NLP research has increasingly shifted from pure algorithmic optimization toward data-centric and system-level design (Zha et al., 2025). Traditional annotation workflows, which rely heavily on manual effort, face severe scalability and cost limitations when constructing large-scale, domain-specific datasets. In this context, annotation tools are expected not only to provide efficient interfaces but also to leverage the zero-shot and few-shot capabilities of LLMs to reshape the data lifecycle. Prior studies show that under extreme low-resource conditions (e.g., 8100 samples), LLMs can substantially outperform conventional supervised models (He et al., 2024), highlighting their potential for high-quality data construction. (Wang et al., 2025)

However, realizing this potential requires more than one-off pre-annotation. Although tools such as BioGraphia (Xu et al., 2025), ITAKE (Song et al., 2024), MEGAnno+ (Kim et al., 2024), and Prodigy (Montani and Honnibal, 2018) have introduced LLM-assisted functionality, most are tailored to specific applications and lack systematic integration of quality control (QC) into annotation workflows. For example, ITAKE emphasizes model monitoring but does not support collaborative annotation or multi-round consistency assessment. Conversely, widely used annotation platforms (e.g., BRAT (Stenetorp et al., 2012), Doccano (Nakayama et al., 2018), TeamTat (Islamaj et al., 2020), LightTag (Perry, 2021)) offer mature project management but provide only isolated consistency statistics, without embedding multi-role collaboration and review into a unified QC pipeline. As a result, existing systems lack continuous, quantitative evaluation of LLM-assisted annotations, preventing a reliable transition from pre-annotation to gold-standard data. In addition, most tools prioritize model-oriented formats (e.g., JSON or Standoff), while real-world data is often managed in Excel, introducing substantial overhead in data preparation and delivery.

To address these challenges, we propose BNLP, a text annotation platform designed for quality-controlled LLM-assisted annotation. Rather than treating LLM outputs as final labels, BNLP explicitly models LLM-generated entities, relations, and attributes as revisable intermediate states within a human-in-the-loop workflow. BNLP integrates multi-role collaboration (Project Manager, Anno-

[*]Equal contribution.
[†]Corresponding author.
[1]Our code is released at https://github.com/BioMedBigDataCenter/BNLP.

Findings of the Association for Computational Linguistics: ACL 2026, pages 23675–23684

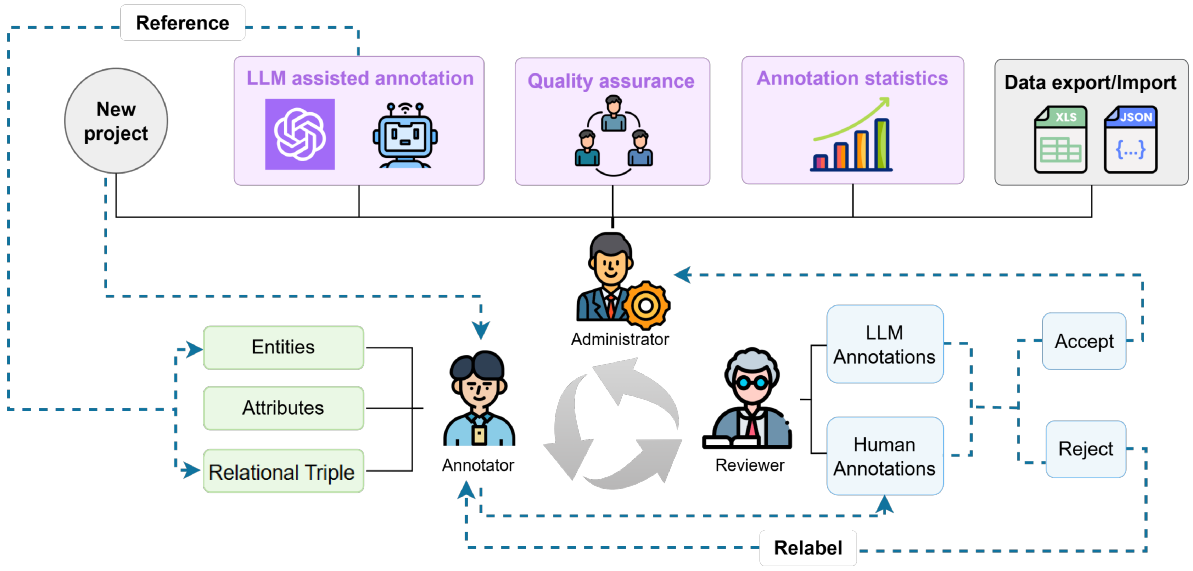

Figure 1: Overview of the BNLP annotation workflow and user roles. The blue dashed lines illustrate the annotation pipeline, spanning project creation, LLM-assisted annotation, manual annotation, review, quality assurance, and data export. The black solid lines represent role-based access control, through which annotators, reviewers, and administrators interact with different system components (entities, attributes, and schemas) and annotation outputs (LLM-generated annotations and human annotations).

tator, Auditor), iterative review cycles, and dynamic visualization of consistency metrics (e.g., Fleiss' Kappa), making quality control an intrinsic part of the annotation process. This design preserves the efficiency gains of LLM pre-annotation while maintainingor even improvingfinal annotation quality. Furthermore, BNLP natively supports AI train-ready formats such as Excel and JSON, ensuring seamless data flow from manual annotation to downstream model training. Our contributions are threefold:

- **Quality-aware LLM-assisted annotation paradigm**: This enables continuous, quantifiable evaluation of LLM outputs through automatic comparison with human revisions.
- **Collaborative data quality control framework**: This integrates multi-role collaboration, multi-round review, and real-time consistency analysis to unlock LLM efficiency under strict quality constraints.
- **AI train-ready data interface**: This supports both Excel and JSON formats to reduce friction across the data lifecycle.

2 System Description

2.1 System Overview

BNLP is a collaborative annotation platform that integrates LLM-assisted labeling into a controllable and evaluable human-in-the-loop workflow for improved efficiency and data quality. It establishes a closed-loop workflow covering automated pre-annotation, manual labeling, auditing, quality control, and statistical analysis. The system organizes annotation around documents, entities, and schemas, supporting multi-level structural annotations such as entities, attributes, and relations.

As shown in Figure 1, the workflow begins with project initialization and progresses through LLM-assisted pre-labeling, manual annotation, auditor revision, quality assurance, and result analysis. The LLM module automatically generates candidate labels for entities and attributes, displayed alongside manual input fields to guide annotators and accelerate labeling.

BNLP supports multi-role collaboration with role-specific permissions. Annotators perform entity, attribute, and relation labeling within a unified interface, supported by UMLS-based concept disambiguation and standardized terminology recommendations. LLM-generated candidates can be selectively adopted or refined. Auditors review annotations and initiate a Relabel request to return tasks for revision when disagreements or quality issues arise. Project managers configure projects, activate LLM and validation modules, and monitor progress through statistical dashboards.

For data reliability, BNLP integrates QA and

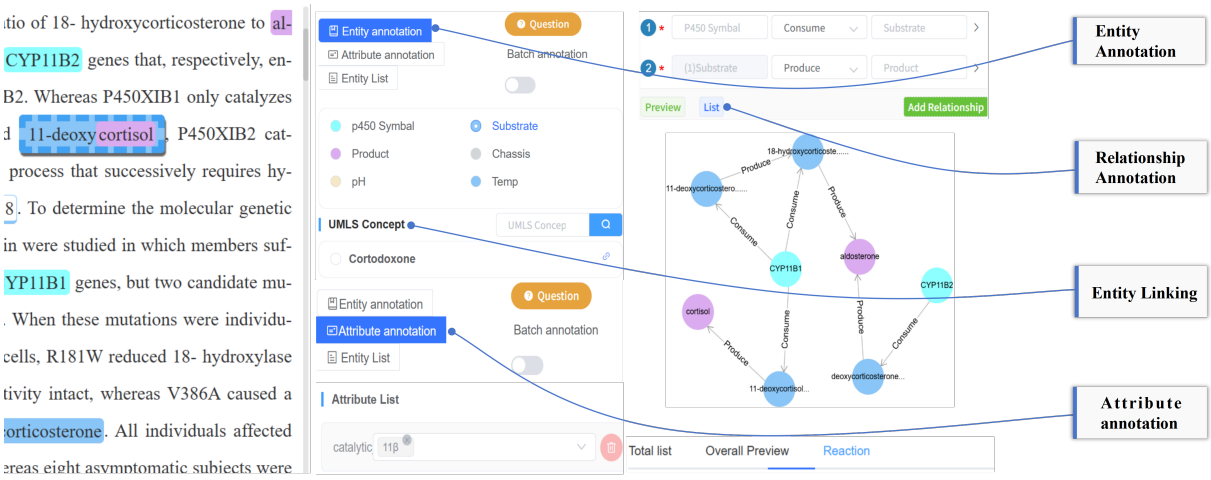

Figure 2: The manual annotation interface of BNLP, highlights support for entity, relation, and attribute annotation, including integrated entity linking functionality and structured visualization of complex biomedical reactions.

annotation statistics modules, supporting inter-annotator agreement metrics such as Fleiss' Kappa and Krippendorff's with visual heatmaps. The platform also supports multi-format data I/O and comparison between automated outputs and expert-validated gold standards, enabling an end-to-end annotation and evaluation cycle.

2.2 Annotation Functions

2.2.1 Manual Annotation

BNLP provides a feature-rich manual annotation environment (Figure 2) built around a unified workflow that supports entity, relation, and attribute labeling within a single interface, reducing context switching and cognitive load. The platform supports continuous and discontinuous entities, as well as overlapping and nested span-scapabilities essential for complex domains like biomedicine but remain in many existing tools. This design enables faithful representation of hierarchical and fragmented semantic structures without forcing premature resolution.

Beyond entity extraction, BNLP natively integrates relation and attribute annotation into the manual workflow. Annotators define directional relations through an interactive graph-based view, where entities are nodes and relations are labeled edges, facilitating both creation and validation. Attribute annotation allows functional or categorical properties to be attached to entities or relations, enabling rich semantic modeling without additional structural complexity. To improve scalability, BNLP introduces a batch annotation mode that applies repetitive labels or attributes across multiple instances simultaneously, reducing redundant manual effort in large-scale corpus construction. BNLP also integrates entity linking into the annotation process, allowing entities to be grounded to external knowledge bases (e.g., UMLS) during labeling rather than as a post-hoc step. This coupling of annotation and normalization distinguishes BNLP from tools treating entity linking as an isolated module and aligns with emerging knowledge-aware annotation practices.

To further improve efficiency, BNLP incorporates LLM-assisted annotation under a human-in-the-loop paradigm (Figure 3). LLMs generate structured pre-annotation candidates rather than final labels, which human annotators then review, revise, or reject. Prompt generation is task-driven and grounded in existing gold-standard data: BNLP automatically constructs prompts including schema definitions, annotation guidelines, and representative few-shot examples, while allowing project managers full control over prompt refinement.

LLM outputs are organized into manageable pre-annotation batches and revised using the same interface as manual annotation, ensuring workflow consistency. BNLP also records execution metadata and resource usage, enabling cost analysis and reproducibility. Pre-annotation results can be exported in standard formats (e.g., JSON, CSV) for downstream quality analysis or model training. Overall, BNLP positions LLM-assisted annotation as a transparent, auditable intermediate stage, establishing a structured foundation for subsequent quality control and project management.

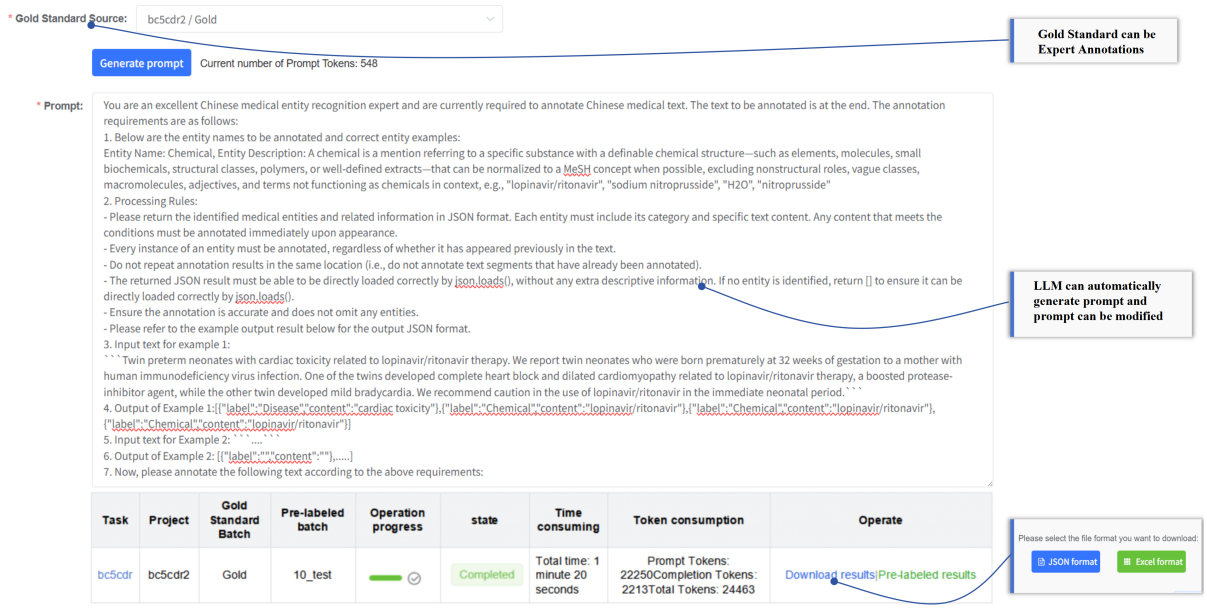

Figure 3: BNLP's LLM-assisted annotation interface supports generating prompts for large language models based on gold-standard data, allows manual prompt editing, and produces structured pre-annotation batches for human annotators to review, correct, and export.

2.2.2 LLM-assisted annotation

To enhance annotation efficiency while preserving data integrity, BNLP integrates Large Language Models (LLMs) as an auxiliary pre-annotation component under a human-in-the-loop paradigm. As shown in Figure 3, LLMs in BNLP do not generate final labels; instead, they produce structured pre-annotation candidates that human annotators must review, revise, or reject. This design reflects the prevailing consensus in recent literature that automated annotation should remain controllable and human-supervised. The LLM-assisted module employs task-driven, dynamic prompt generation.

Project managers can select existing gold-standard datatypically expert-labeled or verified samplesas references for prompt construction. When selecting few-shot examples for LLM prompts, BNLP incorporates an enhanced mechanism using dense retrieval-based methods to identify semantically similar examples, providing the model with more relevant in-context demonstrations. BNLP automatically composes prompt templates that integrate entity definitions, annotation guidelines, and representative few-shot examples, ensuring alignment with project-specific schemas while reducing semantic ambiguity. Full editorial control over prompts enables iterative refinement as annotation guidelines evolve.

For extremely long documents, BNLP employs a preprocessing script that automatically segments input text into smaller chunks fitting within the LLM's context window. Each chunk is processed independently, with final annotations aggregated at the document level. This design avoids silent truncation, ensures complete document coverage, and maintains compatibility with models of varying context window sizes.

After configuration, LLMs generate pre-annotations as discrete batches, managed under the same protocols as manually annotated data. Annotators revise results through identical interfaces and workflows, ensuring seamless integration. Treating pre-annotations as batches supports systematic progress tracking and quality assessment. BNLP records detailed execution metadata and resource consumption during LLM inference, to enable cost analysis and informed project-level decisions. Pre-annotation outputs can be exported in standardized formats (e.g., JSON or CSV) for reuse in downstream quality control, error analysis, or model training.

Overall, BNLP frames LLMs not as replacements for human judgment but as an auditable intermediate step, laying a structured foundation for the project management and quality control described in Section 2.3.

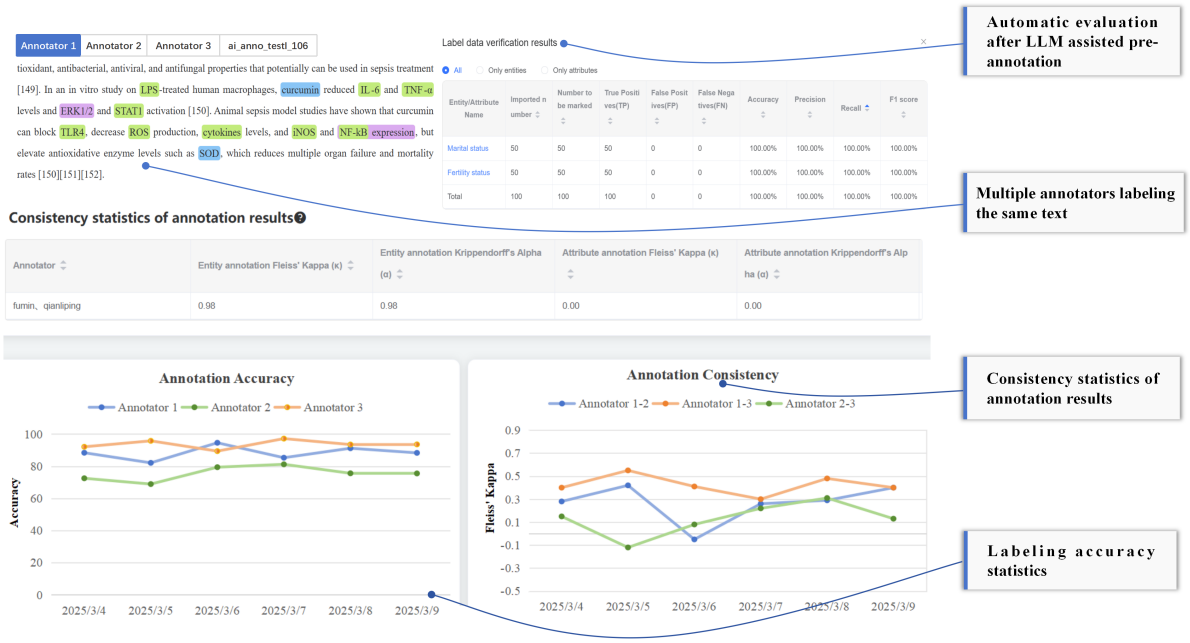

Figure 4: BNLP's project management and quality control features include automatic evaluation of LLM-assisted pre-annotations, support for multiple annotators working on the same text, and reporting of annotation accuracy alongside inter-annotator agreement statisticssuch as Fleiss' Kappa and Krippendorff's Alphathrough tabular and visualized views.

2.3 Project management and quality control

Within the dual-mode annotation frameworkwhere LLM-assisted labeling operates alongside manual effortsBNLP provides a systematic project management and quality control (QC) mechanism. This module, as illustrated in Figure 4, enables continuous monitoring and evaluation of both the annotation process and resulting datasets. By organizing tasks into a hierarchical ProjectBatchAnnotator structure, the system manages data from diverse sources (e.g., purely manual, LLM pre-annotations, and human-revised outputs), ensuring full traceability and reproducibility of the workflow.

Inter-Annotator Agreement (IAA) Analysis. BNLP supports multi-annotator collaboration over identical text spans and automatically computes consistency metrics to quantify annotation quality. The system integrates standard statistical indices, including Fleiss' Kappa and Krippendorff's Alpha, for entity and attribute-level agreement analysis (see Figure 4, middle section). These metrics are generated as cross-sectional snapshots or as dynamic updates synchronized with project progress, enabling managers to identify labeling ambiguities or guideline inconsistencies.

Accuracy Assessment and Validation. Beyond consistency, BNLP offers batch-based accuracy verification. For LLM-generated datasets, the system supports automated comparison against either gold standards or human-revised benchmarks, calculating standard performance metricsincluding Precision, Recall, and F1-score (see Figure 4 upper section)to evaluate the empirical efficacy of model-assisted labeling across varying task configurations. This mechanism transforms the LLM from a "black-box" acceleration tool into an auditable and quantifiable auxiliary component.

Visual Analytics and Lifecycle Monitoring. To capture the evolution of quality over time, BNLP provides visualizations of accuracy and consistency trends (see the lower section of Figure 4). This built-in designincorporating statistical analysis directly into the platformeliminates the need for external evaluation scripts and embeds quality control as an intrinsic part of the annotation lifecycle.

In summary, the project management and QC modules in BNLP synergize with the LLM-assisted mechanism: the project management module provides the evaluation framework, while the LLM mechanism supplies structured, traceable data batches for analysis. This integration ensures efficiency gains achieved through LLMs are balanced by rigorous, quantifiable assessment and long-term maintainability.

Tools	Data format			Annotation Capabilities						Quality control					Usability			
	D1	D2	D3	A1	A2	A3	A4	A5	A6	Q1	Q2	Q3	Q4	Q5	U1	U2	U3	U4
Brat	✓	×	✓	✓	×	✓	✓	×	✓	×	×	✓	×	×	✓	×	×	✓
Doccano	✓	×	✓	✓	×	×	×	×	✓	×	×	✓	×	×	✓	✓	×	✓
TeamTat	✓	×	×	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓	×	✓	✓
Potato	×	×	✓	✓	×	×	×	×	✓	×	×	✓	×	×	✓	✓	✓	✓
YEDDA	×	×	×	✓	×	×	×	×	✓	✓	✓	✓	✓	×	✓	×	×	✓
Prodigy	✓	✓	✓	✓	×	×	×	✓	✓	✓	✓	×	✓	×	×	✓	×	✓
EEVEE	✓	×	✓	✓	×	×	×	×	✓	×	×	×	×	×	✓	✓	✓	✓
MEGAnno	✓	×	×	×	×	×	×	✓	✓	×	×	×	×	×	×	✓	×	×
Doctron	✓	×	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
ITAKE	-	-	-	✓	×	✓	✓	✓	✓	×	×	×	×	✓	-	-	-	✓
BNLP(ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of the functions of various annotation systems.

Note: The following evaluation criteria are defined as follows: "-" indicates the information is unstated or the page is inaccessible, "✓" indicates the feature is supported, and "×" indicates it is not supported. The specific features include: D1 for pre-annotation data upload support; D2 and D3 for Excel and JSON import/export support; A1 for Named Entity Recognition (NER); A2 for Entity Linking with database ID selection or manual disambiguation; A3 and A4 for triplet and attribute definition/annotation; A5 for LLM-assisted annotation; A6 for configurable annotation modes; Q1 for multi-user simultaneous annotation; Q2 for multi-stage review workflows (annotation, review, rejection, and re-annotation); Q3 for multi-user collaboration and management; Q4 for consistency assessment; Q5 for automatic evaluation of LLM-assisted results against a gold standard; U1 for open-source status; U2 for active maintenance and updates; U3 for the availability of an online version; and U4 for multi-language support.

3 Evaluation

3.1 Qualitative Comparison with Existing Tools

A wide range of text annotation systems has been developed, ranging from lightweight manual tools to collaborative and model-assisted platforms. Classic systems such as brat (Stenetorp et al., 2012) and Doccano (Nakayama et al., 2018) support span-based annotation with limited collaboration, while TeamTat (Islamaj et al., 2020) emphasizes project management and consistency statistics. Task-oriented tools including Potato (Pei et al., 2022), YEDDA (Yang et al., 2018), and EEVEE (Sorensen et al., 2024) target specific annotation scenarios but rely on custom or task-specific formats. Prodigy introduces active learningdriven interaction, though its implementation is proprietary. Doctron (Irrera et al., 2025) is designed for information retrieval evaluation with collaborative voting, and recent tools such as MEGAnno+ (Kim et al., 2024) and ITAKE (Song et al., 2024) integrate LLMs primarily to improve automation efficiency.

We compare ten representative systems across four dimensions data format support, annotation expressivity, quality control, and usability (Table 1). Overall, existing tools exhibit fragmented designs, with no single platform jointly supporting complex semantic annotation, collaborative quality assurance, and application-oriented data interfaces. Most systems prioritize downstream model training and thus depend on JSON or task-specific formats (e.g., brat (Stenetorp et al., 2012), TeamTat (Islamaj et al., 2020), YEDDA (Yang et al., 2018)), while offering limited support for human-centric interfaces such as Excel, leading to additional procedural overhead. Moreover, many tools are restricted to basic NER and lack integrated support for entity linking, relation extraction, or attribute annotation. Although brat and TeamTat handle relations, they do not provide unified normalization or attribute modeling. In contrast, BNLP unifies entities, linking, relations, and attributes within a configurable schema, enabling broader task coverage.

More critically, quality control in existing systems is often fragmented or retrospective. Even LLM-enabled platforms primarily emphasize automation rather than systematic evaluation of model outputs. BNLP addresses this gap by integrating multi-role collaboration, multi-stage review, inter-annotator agreement analysis, and LLM-based quality monitoring into a single work-

Tools	Brat	YEDDA	INCEPTION	BNLP (ours)
Annotator 1 (s)	107	73	149	45
Annotator 2 (s)	111	136	138	79
Annotator 3 (s)	178	159	155	142
Annotator 4 (s)	204	184	258	162
Average (AVG $\pm$ STD)	150 $\pm$ 48.55	138 $\pm$ 47.56	175 $\pm$ 55.78	107 $\pm$ 54.40

Table 2: Comparison of annotation efficiency with different annotation tools. STD denotes the sample standard deviation.

flow, making quality assurance intrinsic rather than auxiliary. By reconciling usability with semantic expressivity and collaborative control, BNLP is well suited for constructing high-fidelity datasets in LLM-augmented annotation settings.

3.2 Quantitative Comparison and Efficiency Analysis

We conducted a comparative evaluation between BNLP and three widely-adopted annotation tools: Brat (Stenetorp et al., 2012), INCEpTION (Klie et al., 2018), and YEDDA (Yang et al., 2018). The primary objective was to analyze efficiency differences across these systems during entity annotation tasks.

Experimental Setup. We utilized the training set of the BioCreative V CDR (BC5CDR) corpus (Li et al., 2016) as the data source, from which 10 documents were randomly sampled. Each document contains two types of entities: Chemistry and Disease. Chemical entities are defined as substances with a distinct chemical structure that can be normalized to concepts. Disease entities refer to text mentions representing clear biomedical pathological states, normalizable to one or more concepts. Four graduate students with prior experience in entity labeling participated in the study. To ensure a competitive baseline: YEDDA and INCEpTION utilized their native recommendation or intelligent assistant features, Brat was used in its pure manual annotation mode, BNLP followed the integrated workflow of LLM-based pre-annotation followed by human refinement. To ensure a fair comparison, time measurements only accounted for the active labeling and correction process, excluding non-annotation tasks such as system installation, account creation, and document uploading.

Results and Analysis. The experimental results are summarized in Table 2. BNLP significantly outperformed the baseline tools in terms of average annotation time. As shown in the results, BNLP recorded an average time of 107 seconds, which is markedly lower than Brat (150s), YEDDA (138s), and INCEpTION (175s). Specifically, BNLP achieved time reductions of 28.7%, 22.5%, and 38.9% compared to Brat, YEDDA, and INCEpTION, respectively. These results demonstrate that the integration of an LLM-assisted pre-annotation mechanism effectively alleviates the operational burden of entity recognition and boundary determination. The fact that BNLP maintains a stable temporal advantage while retaining human-in-the-loop verification further validates the practicality and significant potential of LLM-assisted pipelines in specialized biomedical entity annotation scenarios.

4 Case Study

4.1 Food-Medicine Homology Knowledge Graph Construction

In the Food-Medicine Homology (FMH) Antivirus and Anti-bacteria Knowledge Graph project, BNLP served as the core infrastructure for text annotation and corpus management. The goal was to systematically extract structured knowledge from large-scale biomedical and FMH literature, covering 21 fine-grained entity types, attributes, and relations (e.g., compounds, source parts, extraction methods, antimicrobial mechanisms, and clinical effects).

Gold-standard driven prompting. The workflow began with domain experts manually annotating 10 documents to construct a gold-standard seed corpus. BNLP embeds complete gold-standard segments into LLM prompts, automatically retrieving expert-labeled entity mentions as few-shot examples grouped by label. For example, when annotating ingredient entities, expert-defined spans demonstrate precise boundary selection and semantic granularity, enabling the LLM to align more closely with expert annotation styles in complex biomedical contexts.

Semantic normalization and structured relations. To address heterogeneity in chemical and botanical nomenclature, BNLP employed its in-

Metric	Manual	LLM Auxiliary	LLM + Manual
F1	80.75%	80.27%	90.14%
Accuracy	68.31%	67.05%	82.05%
Recall	74.60%	77.63%	84.21%
Average annotation time/person	27.25 min	1.25 min	7 min

Table 3: Comparison of annotation speed and quality between manual annotation and LLM-assisted annotation with human revision. The time required for LLM-assisted annotation comprises task configuration time and LLM inference time.

tegrated UMLS-based concept disambiguation to recommend standardized terms during annotation, reducing conflicts in downstream entity alignment and cross-document merging. Additionally, predefined relation templates (e.g., IngredientMechanismTarget Molecule) allowed annotated data to be directly ingested into the knowledge graph pipeline without further transformation.

Quality control via consistency analysis. To mitigate stylistic divergence in fine-grained labels, Fleiss' Kappa and Krippendorff's Alpha were applied. BNLP's dendrogram-based schema visualization enabled project managers to quickly identify inconsistencies in boundary decisions (e.g., splitting vs. merging antimicrobial descriptions) and recalibrate guidelines. The project ultimately produced 19,536 high-quality structured knowledge fragments, forming a reliable foundation for the FMH knowledge graph.

4.2 P450 Enzyme Metabolic Reaction Annotation

BNLP was applied to annotate P450 enzyme metabolic reactions from 12 PubMed articles, aiming to delineate reaction boundaries and benchmark the reliability of LLM-assisted enzyme metabolism extraction. Annotators first identified core entities, including P450 enzymes, substrates, products, and intermediates. BNLP's high-precision highlighting facilitated rapid localization within long and fragmented experimental narratives. For metabolic parameters (e.g., K_m, V_{max}), the attribute annotation interface enabled direct binding between entities and their properties. BNLP also supports cell-level tabular annotation via uploaded experimental tables, unifying structured tables and textual evidence within a single corpus. Given the multi-step nature of P450 metabolism, BNLP's relation annotation module was used to construct substrateenzymeproduct triplets through an interactive graph interface. Dispersed reaction descriptions were modeled as ordered sequences of linked triplets, yielding a clear hierarchical representation of complex metabolic pathways.

After initial annotation, expert auditors refined the outputs using BNLP's review interface. Specifically, two domain experts with prior experience in enzyme metabolism annotation participated in this study. Both reviewers have backgrounds in pharmacology and biomedical literature curation. They independently reviewed the annotations and resolved discrepancies through discussion to ensure the consistency of the final labels. They were responsible for verifying the correctness of entity boundaries, relational structures, and attributes prior to the final integration of the corpus. Annotation efficiency was evaluated on 20 document segments (3,914 words) by comparing manual annotation, LLM-only pre-annotation, and the "LLM + Human-in-the-loop" workflow against P450RDB (Zhang et al., 2024). BNLP utilizes the standard OpenAI API protocol for LLM calls, supporting seamless switching and the configuration of custom system parameters. We report performance using DeepSeek-V3 (set to $temperature = 0.1$ and $max\_tokens = 16,000$). If current state-of-the-art closed-source models, such as GPT-5.3 Pro, were employed, performance would likely be further enhanced. As shown in Table 3, LLM pre-annotation reduced annotation time by 26 minutes (95.4% speedup) compared to the manual baseline, but exhibited inferior quality when used alone. Incorporating human revision preserved substantial efficiency gains reducing time by 20.25 minutes (74.3% speedup), while improving annotation quality by 11.6% in F1-score.

Overall, this case study produced a high-fidelity P450 metabolic reaction corpus with fine-grained metabolic annotations. BNLP's unified support for entities, attributes, relations, and tabular data enables reliable modeling of complex biochemical semantics, providing a strong foundation for downstream LLM fine-tuning, reaction boundary refinement, and model reliability evaluation. The results demonstrate BNLP's effectiveness in spe-

cialized domains and validate its humanLLM collaborative annotation paradigm.

5 Conclusion

We present BNLP, a natural language annotation platform designed for controllable quality in LLM-assisted annotation, particularly for low-resource and expert-dependent scenarios. Unlike existing approaches that treat LLMs as one-off labeling tools, BNLP models LLM outputs as revisable intermediate states and embeds them into a collaborative workflow with multi-role coordination, and consistency analysis, making quality control intrinsic to the annotation process. BNLP integrates inter-annotator agreement with automated comparisons between LLM outputs and human revisions, enabling continuous quantifiable quality monitoring. Under these constraints, the platform effectively harnesses LLM efficiency for entity, relation, and attribute annotation while substantially reducing manual effort. Experiments show that BNLP significantly accelerates annotation while maintaining or improving data quality, validating the effectiveness of our humanLLM collaborative annotation paradigm.ă

Limitations

Despite its advantages, BNLP has several limitations. First, the current evaluation focuses on the biomedical domain (e.g., BC5CDR and P450), which may limit empirical demonstration of cross-domain generalization. We chose biomedical datasets because they represent a challenging, terminology-dense, and resource-scarce setting where annotation quality control is particularly important. However, BNLP itself is domain-agnostic: the platform operates on user-defined schemas and does not rely on domain-specific heuristics. Second, although BNLP supports a configurable offline "AnnotationTrainingEvaluation" loop, this process is not yet tightly integrated into the interactive annotation workflow. While this design prioritizes flexibility and user control, it limits rapid model adaptation to newly annotated data.Future efforts will focus on finer-grained training triggers and semi-automated update mechanisms to accelerate model evolution while preserving human oversight.

Acknowledgments

We thank all anonymous reviewers. This work was supported by the National Key R&D Program of China (Grant Nos. 2025YFF0511500 and 2023YFA0915500), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA0460203), and the Shanghai Science and Technology Innovation Action Plan (Grant No. 23JS1401500). We thank Ruijin Luo and his team from Shanghai South Gene Technology Co., Ltd. for their support in website development.

References

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Ornella Irrera, Stefano Marchesin, Farzad Shami, and Gianmaria Silvello. 2025. Doctron: A web-based collaborative annotation tool for ground truth creation in ir. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3488–3497.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. INCePTiON: A high-level platform for iterative corpus annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus:

a resource for chemical disease relation extraction. *Database*, 2016.

Ines Montani and Matthew Honnibal. 2018. *Prodigy: A modern annotation tool for AI, Machine Learning and NLP*. Explosion AI, Berlin, Germany.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Tal Perry. 2021. Lighttag: Text annotation platform. In *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations*, pages 20–27.

Jiahe Song, Hongxin Ding, Zhiyuan Wang, Yongxin Xu, Yasha Wang, and Junfeng Zhao. 2024. ITAKE: Interactive unstructured text annotation and knowledge extraction system with LLMs and ModelOps. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 326–334, Bangkok, Thailand. Association for Computational Linguistics.

Axel Sorensen, Siyao Peng, Barbara Plank, and Rob Van Der Goot. 2024. Eevee: An easy annotation tool for natural language processing. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 216–221.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Xi Xu, Sumin Jo, Adam Officer, Angela Chen, Yufei Huang, and Lei Li. 2025. BioGraphia: A LLM-assisted biological pathway graph annotation platform. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 480–486, Suzhou, China. Association for Computational Linguistics.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A lightweight collaborative text annotation tool. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42.

Yang Zhang, Xianrun Pan, Tianyu Shi, Zhifeng Gu, Zhaochang Yang, Minghao Liu, Yi Xu, Yu Yang, Liping Ren, Xiaoming Song, and 1 others. 2024. P450rdb: A manually curated database of reactions catalyzed by cytochrome p450 enzymes. *Journal of advanced research*, 63:35–42.