Constructing CCEE an LLM evaluation dataset for Complex Context-aware Event Extraction for gene regulatory networks

Frederik Labonté[1,2], Lucie Flek[1,2]

[1]b-it/University of Bonn
[2]Lamarr Institute for ML and AI

Abstract

This paper presents a first look at CCEE (Complex Context-aware Event Extraction), a currently in the works novel evaluation dataset for context-rich gene regulatory network extraction from scientific literature. We propose an annotation scheme for cancer research papers, capturing both core gene interactions and extensive contextual information across 10-14 categories per event, addressing limitations in existing datasets to test construction of disease specific biomedical knowledge graphs. Unlike previous datasets that focus primarily on entity connections of isolated triplets, CCEE links contextual attributes directly to gene regulatory events, providing a more integrated representation of scientific knowledge. We illustrate the annotation on 9 papers manually labeled by multiple experts, and give a first impression of challenges and ways to address them. Additionally we show first evaluations of LLMs as an annotation system. While it under performs human experts in interaction type labeling, it matches human performance on attributing entities as context to interactions.

Keywords

Natural Language Processing, Gene Regulatory Networks, Text Annotations, Knowledge Graphs, Cancer

1. Introduction

The gene regulatory networks (GRNs) underlying diseases like cancer can be represented through knowledge graphs (KGs). Which then can be used to find therapeutic targets [1, 2]. GRNs in particular are context dependent and highly influenced by external factors [3, 4], making the extraction of experimental conditions a crucial addition to the entities and relations involved. Due to data scarcity The automated construction of GRNs remains challenging [5], even more so the extraction of regulatory events and their additional biological context. This is due to data scarcity and focus of existing datasets on the direct relations between entities rather than the surrounding context, e.g. experimental conditions. The data sets that capture connections individually can make linking them to the regulatory event ambiguous.

To address this, we focus on linking contextual information directly to interaction triplets. Here we give a first look into the work in progress on the evaluation dataset for Complex Context-aware Event Extraction **CCEE**. Together with biomedical experts, we annotated 9 full-text articles on cancer, identifying more than 200 matched events, comprising of the core interaction triplet and additional 9 or 13 contextual categories (**Table 1**).We choose cancer as the domain, since the changes in GRNs often lead to the disease and the amount of publications is so vast that even human experts can not keep up with all new findings. Making it a perfect candidate for automated KG construction. Providing longer more in depth events than typical triplets.

Since LLMs offer a potential solution for fields lacking extensive training data, including biomedical annotations [6, 7], we test the performance of Mistral Large 2 [8], providing a preliminary qualitative analysis to analyze which categories of LLM annotations are effective.

8th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, 2025, June 1, Portorož, Slovenia
flabonte@uni-bonn.de (F. Labonté); flek@bit.uni-bonn.de (L. Flek)
0000-0002-5995-8454 (L. Flek)

Related Work

Multiple datasets have been developed for biomedical event extraction, each with distinct focus and annotation approach. Examples include MLEE [9], GE11 [10], CG [11], and BioRED [12].

The **MLEE (Multi-Level Event Extraction)** corpus contains 262 PubMed abstracts on angiogenesis, with $\tilde{8}$,000 entity mentions and 6,000 events. It spans 19 entity types and 40 event types across molecular to organism levels, capturing multi-scale biological interactions.

The **GE11 (GENIA Event)** corpus, developed for BioNLP Shared Task 2011, includes 1,200 MEDLINE abstracts annotated for gene expression, protein interactions, and biochemical events, serving as a benchmark for molecular event extraction.

The **CG (Cancer Genetics)** corpus, developed for BioNLP Shared Task 2013, contains 600 documents with more than 17,000 events, 21,683 entity annotations, and 917 relations, focusing on cancer-related pathological and physiological processes.

The **BioRED (Biomedical Relation Extraction Dataset)** spans 600 PubMed abstracts, annotating gene, disease, and chemical relations. It uniquely distinguishes novel findings from established knowledge, enhancing relation extraction capabilities.

While these datasets provide a significant basis for biomedical information extraction, they share several limitations. With the exception of BioRED, these datasets all focus on abstracts, potentially leaving out additional information about experimental design. Additionally, none of them focus on capturing contextual information around annotated events. While BioRED also captures disease, species, cell line and chemical category, it does not annotate non-chemical treatments, the level of regulation, the tissue or general associated factors.

2. Dataset Creation and Annotation Methodology

In contrast to the existing resources, **CCEE**, was designed for the purposes of applications that require a deeper understanding of the context and relationships within biomedical research, like disease KGs and specifically to test generative models as information extractors. The events we extract are enriched with contextual data spanning 13 context categories for G-D and 9 for G-G interaction 1. This allows us to test the capability of models to detect and extract these events which build the basis for KG construction and automated context aware knowledge linking. The 3 main differentiating factors from existing datasets are the following.

1. **Full-text analysis:** Unlike the abstract-focused approaches of MLEE, GE11, and PHEE, our dataset annotates complete research papers, capturing the rich context typically found in methods and results sections that is absent from abstracts.
2. **Comprehensive contextual integration:** We are capturing two main event types Gene-Gene (G-G) and Gene-Disease (G-D) interactions, adding extensive contextual information to each event, to better reflect the contextual nature of GRNs. Our annotation schema connects direct regulatory relationships with their broader experimental and biological context(Table 1).
3. **Event-centric contextual linking:** Rather than treating disease states, experimental models, and treatments as separate events, our approach explicitly links these elements as contextual attributes of gene regulatory events, creating a more integrated representation of the scientific knowledge.

To select our annotation corpus, we filtered the PubMed Open Access database with a specific focus on cancer research literature. Documents were pre-processed using Hunflair 2 [13] to identify gene, species, and disease entities, which enabled us to filter papers based on gene frequency and cancer mentions. to avoid outliers and to avoid annotating papers with no gene specific information. We

removed the upper and lower 25% of papers by unique gene count, we randomly selected documents for annotation.

To optimize the annotation process while preserving contextual integrity, we implemented a windowed information extraction. For each gene identified by Hunflair 2, we extracted the surrounding sentence in both directions, and merged overlapping windows. Splitting each paper on average into 11.5 blocks containing 11.43 sentences, each block got its own ID. This strategy maintained critical contextual information while reducing annotator cognitive load.

Our annotation schema was then developed iteratively through consultation with expert medical researchers who specialize in lung cancer and radiation treatment research. Due to the absence of established guidelines for complex event annotation with multiple contextual categories, we are refining our approach through successive iterations. We recruited two annotators with formal education in biology and specific experience in genetics and molecular biology. The annotators underwent initial training through shared annotation exercises before proceeding to independent annotation work.

The current annotation schema distinguishes between two primary event types: G-G and G-D relationships. When annotators identified these relationships, they systematically documented the connection and populated multiple contextual categories those can be seen in table 1. We use 3 types of labels, predefined labels, and free text fields which are used to capture hard to define categories like associated factors. Those free text fields are also included to specifically test LLM performance on understanding potentially important information that doesn't fit the annotation scheme.

2.1. Annotation Matching Methodology

Establishing a reliable matching procedure for evaluating inter-annotator agreement presents significant challenges in relation extraction tasks, particularly when annotations lack unique identifiers. In our dataset, annotations are only associated with source sentences and block IDs, neither of which provides sufficiently granular identification as multiple distinct events may originate from the same textual source.

To address this challenge, we implemented a context-based matching approach. For G-D relationships, two annotations were considered a match when both the disease entity and gene entity overlapped between annotators within the same block ID. Similarly, for G-G relationships, annotations were matched when both the primary gene and connected gene aligned between annotators within the same block ID.

Potentially, more stringent matching criteria could further increase precision by incorporating additional contextual elements. For instance, in cases where an event appears multiple times with differing contextual information (e.g., the same G-G interaction mentioned in relation to different cancer types), our current approach may produce ambiguous matches. However, this would reduce the number of categories available for annotation scheme evaluation, as categories used for matching were excluded for inter-annotator agreement.

2.2. F1 Score Calculation

For each annotation, we determine if a corresponding one existed in the same assigned chunk from the second annotator. The total number of unmatched entries per annotator was subtracted from their total annotation count to derive true positives. Unmatched annotations formed false positives and negatives.This approach yielded asymmetrical results due to the fuzzy entity boundaries[14]. Results are reported in Table 2.

Table 3 further details the matching statistics, indicating that approximately 61.7% of Anno2's G-G relationships were matched with Anno1, while 61.1% of Anno1's G-G relationships were matched with Anno2. For G-D relationships, the matching rates were 54.1% and 61.2% for Anno2 and Anno1, respectively.

While F1 scores provide insights into annotation overlap, more robust statistical measures are required to assess true inter-annotator agreement while accounting for chance agreement. Therefore,

we extended our evaluation using Cohen's Kappa and Gwet's AC1 coefficients [15].

2.3. LLM-Assisted Annotation Matching

A significant methodological challenge arose from our use of unnormalized entities and free-text categorical annotations. Traditional string-matching approaches are inadequate for such data, as

Table 1
Overview of the Categories and type of annotation done respectively.

Category	Annotated for	Short explanation	Label
Associated_Factors	G-G, G-D	Free text, keywords, describing additional info	text
Cellline	G-G, G-D	If a cell-line was mentioned	entity
Confidence	G-G, G-D	perceived confidence of an event	predefined
Connected_Gene_Protein	G-G	A gene connected to another gene	entity
Connection_Disease	G-D	role of the gene connected to the disease (progression, suppression, connected)	predefined
Connection_Treatment	G-D	Effect of the treatment	predefined
Connection_Type	G-G	Lable to describe connection between gens	predefined
Disease	G-G, G-D	The disease that is mentioned in connection to a gene	entity
Disease_Mechanism	G-D	The mechanisms the gene is involved in that effect the disease	text
Gene_Protein	G-G, G-D	The Gene that is mentioned in connection to a disease or another protein	entity
Regulation_Level	G-G	In which level of gene regulation does the gene gene interaction take place	predefined
Sentence	G-G, G-D	Where in the text did we find the supporting evidence	number
Species	G-G, G-D	The species the connection was observed in	entity
Tissue	G-G, G-D	The Tissue the connection was observed in	entity
Treatment_Exposure	G-G, G-D	What treatments were used e.g. drug tests	
Treatment_Mechanism	G-D	What does this treatment effect	text
Type_of_Evidence	G-G, G-D	What kind of evidence do we have direct, indirect, citation	predefined
connection_disease_marker	G-D	If the gene is a marker for the disease or its outcomes	predefined
treatment_mechanism_type	G-D	If the treatment is acting as a transporter or through targeting	predefined

Table 2
Resulting asymmetrical scores between annotator 1 and 2.

Relationship Type	Precision	Recall	F1 Score
G-G Anno1	0.61	0.53	0.57
G-D Anno1	0.61	0.50	0.55
G-G Anno2	0.62	0.71	0.66
G-D Anno2	0.54	0.66	0.60

Table 3
Details of matching statistics. G-G Anno1 e.g. refers to the percentage of the annotations of annotator 1 accepted by matching to annotator 2.

Relationship Type	Total Anno1	Matched (%)	Unatched (%)	Total Anno2	Matched (%)	Unmatched (%)
G-G	201.0	61.7	38.3	175.0	61.1	38.9
G-D	209.0	54.1	45.9	170.0	61.2	38.8

they fail to capture semantic equivalence when annotators use different phrasing to express identical concepts. To address this limitation, we implemented a LLM as a judge approach for determining categorical matches. For multi-value categories, we considered annotations to match if at least one entry overlapped semantically as determined by the LLM, the prompt can be found in the project GitHub. To validate this approach, a subset of categories was also matched manually by human annotators. This validation process revealed that the LLM-based matching achieved consistently high agreement with human judgment, with Cohen's Kappa above 0.86 across tested categories (See Appendix), while marginally under performing rules-based matching in categories where predefined labels are meant to be matched.

2.4. Sentence-Level Agreement Analysis

Beyond entity and relationship matching, we also evaluated the consistency of text selection by examining the specific sentences from which annotators extracted information. This analysis reveals an interesting pattern: while annotators achieved only moderate rates of exact sentence matching (38.9% for G-G and 24.7% for G-D relationships), the average matching rate increases to 57% and 51%, respectively. This agreement improves substantially when considering a window of ±1 sentence around the selection. The average range match, which calculates the overlap percentage while including adjacent sentences, reaches 78.8% for G-G and 83.0% for G-D relationships. The higher range-match percentage for G-D relationships, despite lower exact matching, suggests that disease-related information is often distributed across adjacent sentences, requiring broader contextual integration.

2.5. Comparison of Agreement Metrics

Figure 1 illustrates the comparative performance of Cohen's Kappa and Gwet's AC1 coefficients across annotation categories for G-G connections. A notable divergence between these metrics is observed in several categories, which can be attributed to the prevalence of unbalanced label distributions in our dataset. [16]

Categories such as Cell Line, Species, and Tissue frequently exhibit this imbalance due to the absence of explicit information in the source text, resulting in numerous empty annotations. Similarly, the Confidence category shows distortion due to a predominance of "high" confidence annotations. This pattern aligns with known limitations of Cohen's Kappa, which can produce paradoxically low values despite high observed agreement when marginal probabilities are imbalanced—an effect often referred to as the "Kappa paradox" in the literature[16].

Gwet's AC1, while conceptually similar to Cohen's Kappa, incorporates adjustments that provide greater stability when analyzing unbalanced categories. In the context of annotation evaluation, a threshold of 0.6 is considered acceptable for downstream applications [17].

Of particular note, the Species category in G-G interactions falls below this threshold with a Gwet AC1 score of approximately 0.54. This lower agreement stems from inconsistent understandings when the context was clearly linked, additionally the use of a very rudimentary annotation environment increases the chance of simple overlooking an entity mention.

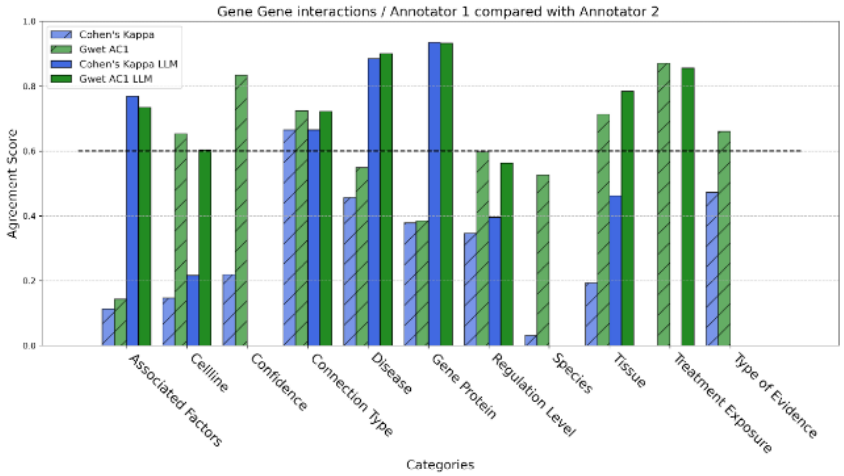

Figure 1: G-G connection compared between the two annotators comparing the kappa scores and gwet scores with and without LLM matching instead of standard string matching.

2.6. Gene-Disease Relationship Agreement Analysis

G-D relationships showed lower inter-annotator agreement than G-G relationships, requiring refinement of annotation guidelines. "Connection Disease" and "Connection Treatment" failed to meet the 0.6 threshold for reliable biomedical annotations [17], indicating interpretation challenges. Annotator discussions revealed issues stemming from category underdefinition. For example, when two genes interact in relation to a specific cancer, it remained unclear whether both genes should be annotated as connected to the disease, or only one—and if so, which one. The definition of when a gene or treatment is considered to regulate a disease also needs clarification. For "Connection Disease," annotators frequently disagreed on directionality, with one assigning directional relationships while the other chose non-directional labels. In "Connection Treatment," 13% of cases had mismatched directional labels, and 23% disagreed on relationship existence, further suggesting definitional problems. Categories like "Species," "Tissue," and "Cell Line" also failed to reach acceptable agreement. The consistent low agreement in G-D relationships, despite acceptable G-G relationship agreement, highlights the greater challenges in annotating gene-disease interactions, driven by less obvious interaction definitions and by extension less clarity of which context to include.

3. Comprehensive Annotation Statistics and Agreement Analysis

3.1. Annotation Volume and Matching Overview

Our annotation corpus comprises a substantial body of biomedical relation extraction annotations distributed across two major event types. Annotator 1 (A1) produced a total of 410 annotated events, while Annotator 2 (A2) generated 345 events. These annotations were distributed relatively evenly between G-G and G-D relationships, with A1 creating 201 G-G and 209 G-D events, and A2 producing 175 G-G and 170 G-D events. The matching analysis revealed that for A1, 124 G-G events and 113 G-D events had at least one corresponding match in A2's annotations. Conversely, 107 G-G events and 103 G-D events from A2 had at least one match in A1's dataset. This yields overall matching rates of approximately 62% for A1's annotations and 61% for A2's annotations, consistent with our earlier F1 score analysis.

3.2. Category-Specific Agreement for G-G Relations

For G-G relationships, we calculated observed agreement across multiple contextual categories. Table 4 presents these agreement scores. These agreement rates reveal substantial variability in annotator

Table 4
Observed agreement rates across categories for G-G relationships.

Category	Observed Agreement
Associated Factors	0.53
Cell Line	0.88
Confidence	0.71
Connection Type	0.59
Disease	0.71
Regulation Level	0.56
Species	0.61
Tissue	0.82
Treatment Exposure	0.91
Type of Evidence	0.77

consensus across different contextual dimensions. Particularly strong agreement was observed for Treatment Exposure (0.91), Cell Line (0.88), and Tissue (0.82), suggesting these categories have well-defined boundaries that facilitate consistent annotation. Conversely, Associated Factors (0.53), Regulation Level (0.56), and Connection Type (0.59) demonstrated marginal agreement levels, indicating potential areas for guideline refinement.

3.3. Annotation Complexity and Volume

The annotation schema implemented in this study captures substantial contextual information beyond the core entity relationships. For G-D annotations, each event required documentation of the gene, disease, and 14 additional contextual attributes, resulting in 3,344 total data points from Annotator 1 and 2,720 from Annotator 2. Similarly, G-G annotations captured the primary gene, connected gene, and 10 contextual attributes, yielding 2,412 data points from Annotator 1 and 2,100 from Annotator 2.

3.4. Annotation Methodological Approach

Our annotation framework incorporated three distinct annotation methodologies to capture the full spectrum of biomedical information:

1. **Named entities:** Capturing specific biomedical entities such as genes, diseases, and treatments
2. **Relations:** Documenting the connections between identified entities
3. **Free text fields:** Providing flexibility to capture contextual information that does not conform to standardized categories

The inclusion of free text fields is essential for preserving information that might otherwise be lost in a more rigid annotation framework, though this approach introduced additional challenges for inter-annotator agreement assessment as discussed in our methodological section.

3.5. Unique Features of the Dataset

The goal of the finished dataset is to advances biomedical NLP by addressing gaps in context-aware gene regulatory network extraction. It captures the complex contextual dimensions of G-G and G-D relationships, providing depth for evaluating automated knowledge graph construction. With annotations from 9 cancer research papers, it includes over 200 matched events and 2,000 contextual data points. Despite lower inter-annotator agreement in G-D relationships, this variability serves as a valuable benchmark for measuring annotation precision and recall.

The dataset combines structured categorical annotations and free-text fields, capturing nuanced contextual information. This design is ideal for evaluating large language models and human annotators. By offering detailed performance metrics across multiple annotation categories, the dataset will help

identify strengths and weaknesses in generative models, providing insights into the challenges of extracting biological context for precision medicine applications.

4. LLM Baseline Experiments

To establish a technological baseline for our complex biomedical relation extraction task. We employed Mistral Large 2, providing it with the same annotation instructions and two representative examples before tasking it with independent annotation to then be compared against human baseline.

Table 5
Performance comparison between Mistral Large and human annotators (A1 and A2)

Annotation Type	Precision (Human as Reference)	Recall (Human as Reference)	F1 Score (Human as Reference)	Precision (AI as Reference)	Recall (AI as Reference)	F1 Score (AI as Reference)	F1 Score human baseline
G-G vs. A1	0.63	0.33	0.43	0.30	0.58	0.39	0.57
G-D vs. A1	0.91	0.43	0.58	0.41	0.86	0.55	0.55
G-G vs. A2	0.74	0.44	0.55	0.42	0.70	0.52	0.66
G-D vs. A2	0.74	0.43	0.54	0.41	0.72	0.53	0.60

The F1 scores are slightly below human inter-annotator agreement, except for G-D annotations compared to Annotator 1, where the LLM performed similarly. Notably, there's a significant asymmetry between precision and recall metrics.

This asymmetry is due to the volume of annotations. Mistral 2 generated more G-D and G-G events than human annotators, but only 30-40% of LLM-annotated events matched human annotations, compared to 55-61% agreement between human annotators.

Sentence selection patterns of both annotators and the LLM showed similar performance.

These findings support that while the LLM may generate more false positives, they can detect many of the events.

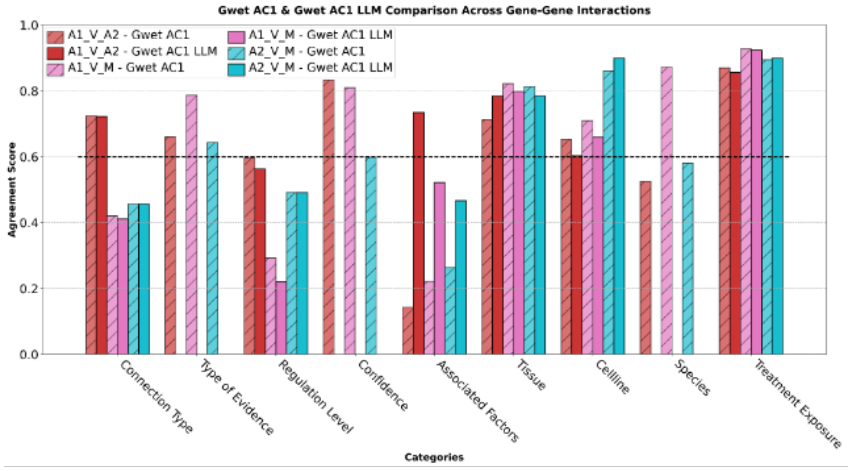

Figure 2: Overview of Annotator compared with mistral across the annotated categories for G-G interactions.

As illustrated in Figure 2 for G-G data, category-specific performance reveals interesting patterns. While Mistral underperformed in categories requiring relationship interpretation between entities, it achieved comparable or superior performance to human annotators in entity extraction categories , for these entity-based categories, the LLM demonstrated higher agreement with both human annotators than the human annotators achieved with each other, suggesting particular strength in expanding events with additional contextual information. A similar pattern was observed in G-D.

5. Conclusion

Our work-in-progress dataset aims to serves as seed and evaluation data for assessing LLM annotation performance and establishing groundwork for synthetic data generation. A key application is testing components for automatic KG construction, specifically relation extraction with contextual categories that enable cross-source data connections.

Preliminary evaluations using Mistral Large 2 showed the model performing below human annotators in interaction labeling, yet achieving comparable or superior performance in attributing context to detected events.

5.1. Limitations & Future Work

This initial exploration of Complex Context-aware Event Extraction in cancer research has limitations due to dataset size and using only two annotators. Inter-annotator agreement challenges, particularly in categories like "Connection Disease & species," revealed guideline ambiguities and annotation environment problems.

We identified three main issues: (1) underdefined criteria for gene-disease events causing disagreement, (2) lack of automated highlighting leading to missed entities, and (3) pre-annotated entities without proper ID linking creating matching problems alongside limited positional data.

To address these challenges, we will: define clear G-D connection criteria similar to BioRED [12] guidelines; redo annotations using software like TeamTat to resolve highlighting, position information, and normalization issues; and extend guidelines with more examples and case differentiation. Future plans include dataset expansion and adding a third annotator for the final release.

We hope to stimulate discussion on integrating contextual information for KG construction and evaluating LLMs as annotators for disease-KG construction.

6. Data Availability and Ethical Considerations

The data set, prompts and the (revised) annotation guidelines will be made available at:
https://github.com/FMLabonte/CCEE_for_GRNs


```
\begin{acknowledgments}
We thank:
The TRA grant of the university of Bonn making this research possible.
The lamarr institute for the ample opportunities for exchange of ideas.
The DLR and Dr. Christine Hellweg for feedback on the usefulness of
 ↪  annotations.
All the people that helped with formatting, feedback and spell checking. Luna
 ↪  Meyer, Valerie Dang, Max Brauner, Liliane Hanfeld and Lukas Grönwoldt.
Georgina Kowalski for the additional annotation work.
\end{acknowledgments}
```


A. Appendix

A.1. LLM as a judge evaluation

Notably, in several instances, the LLM demonstrated higher agreement with individual human annotators than the human annotators achieved with each other, suggesting the LLM's effectiveness as an impartial judge for this task.

For this matching procedure, we employed GPT-4o with a temperature setting of 0 to maximize deterministic outputs. This approach was primarily applied to categories containing full-text or entity annotations where semantic rather than exact matching was appropriate.

A.2. Tables

Table 6
Cohen's Kappa scores between annotators and the LLM judge across various annotation categories.

Category	Anno1 vs. GPT	Anno1 vs. Anno2	Anno2 vs. GPT
Treatment	0.99	-	-
Associated Factors	0.86	-	-
Disease	0.87	0.822	0.89
Cell Line	1.00	-	-
Treatment Mechanism	0.93	-	-
Species	1.00	-	-
Disease Mechanism	0.93	0.81	0.98
Tissue	0.95	-	-

Table 7
Sentence-Level Agreement Statistics for Human Annotators.

Relationship Type	Total Matched Events	Exact Sentence Match(%)	Average Sentence Overlap (%)	Average Range Match (± 1 sentence) (%)
G-G	131	51 (38.9%)	57.0%	78.8%
G-D	146	36 (24.7%)	50.7%	83.0%

Table 8
Sentence Selection Comparison Between Human Annotators and LLM

Comparison	Relationship Type	Total Matched Events	Exact Sentence Match (%)	Average Sentence Overlap (%)	Average Range Match (± 1 sentence) (%)
Human vs. LLM	G-G	164	65 (39.6%)	60.1%	81.5%
Human vs. LLM	G-D	223	90 (40.4%)	60.5%	99.2%
Anno1 vs. LLM	G-G	121	47 (38.8%)	59.2%	81.1%
Anno1 vs. LLM	G-D	235	59 (25.1%)	46.9%	81.4%

Table 9
Detailed Matching Statistics Between Human and LLM Annotations

Annotation Type	Human Total	Human Matched (%)	Human Unmatched (%)	LLM Total	LLM Matched (%)	LLM Unmatched (%)
G-G (Anno1)	201	149 (74.1%)	52 (25.9%)	339	141 (41.6%)	198 (58.4%)
G-D (Anno1)	209	155 (74.2%)	54 (25.8%)	362	150 (41.4%)	212 (58.6%)
G-G (Anno2)	175	111 (63.4%)	64 (36.6%)	339	101 (29.8%)	238 (70.2%)
G-D (Anno2)	170	154 (90.6%)	16 (9.4%)	362	147 (40.6%)	215 (59.4%)

A.3. Figures

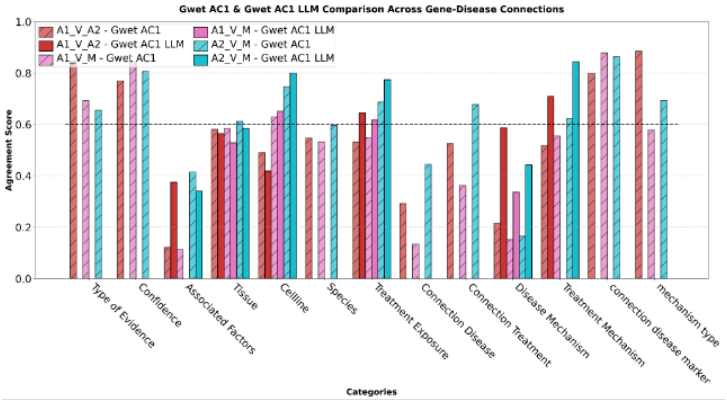

Figure 3: Overview of Annotator compared with mistral across the annotated categories for G-D interactions.

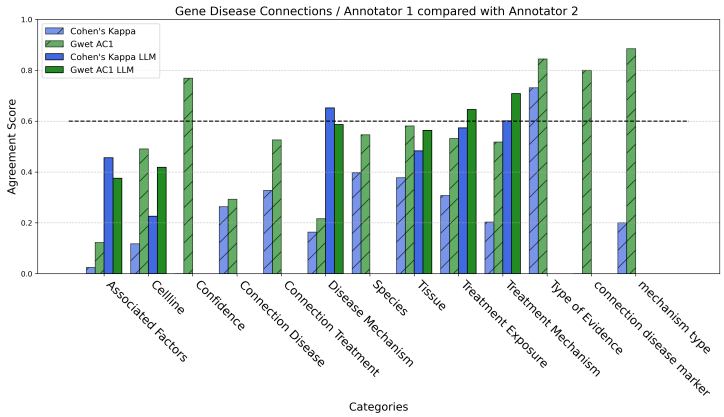

Figure 4: G-D connection compared between the two annotators comparing the kappa scores and gwet scores.

Declaration on Generative AI

During the preparation of this work, the author(s) used Mistral and Claude Sonnet 3.5 for grammar and spelling checking. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. B. Madhamshettiwar, S. R. Maetschke, M. J. Davis, A. Reverter, M. A. Ragan, Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets, Genome Medicine 4 (2012) 41. URL: https://doi.org/10.1186/gm340. doi:10.1186/gm340.

[2] T. Raschka, M. Sood, B. Schultz, A. Altay, C. Ebeling, H. Fröhlich, Ai reveals insights into link between cd33 and cognitive impairment in alzheimer's disease, PLOS Computational Biology 19 (2023) 1–24. URL: https://doi.org/10.1371/journal.pcbi.1009894. doi:10.1371/journal.pcbi.1009894.

[3] F. Emmert-Streib, M. Dehmer, B. Haibe-Kains, Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks, Frontiers in Cell and Developmental Biology 2 (2014) 38. URL: https://doi.org/10.3389/fcell.2014.00038. doi:10.3389/fcell.2014.00038.

[4] Y.-H. Tu, H.-F. Juan, H.-C. Huang, Context-dependent gene regulatory network reveals regulation dynamics and cell trajectories using unspliced transcripts, Briefings in Bioinformatics 24 (2023) bbac633. URL: https://doi.org/10.1093/bib/bbac633. doi:10.1093/bib/bbac633.

[5] S. Geißler, The Kairntech Sherpa – an ML platform and API for the enrichment of (not only) scientific content, in: G. Rehm, K. Bontcheva, K. Choukri, J. Hajič, S. Piperidis, A. Vasiļjevs (Eds.), Proceedings of the 1st International Workshop on Language Technology Platforms, European Language Resources Association, Marseille, France, 2020, pp. 54–58. URL: https://aclanthology.org/2020.iwltp-1.9/.

[6] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, J. Steiner, I. Laish, A. Feder, Llms accelerate annotation for medical information extraction, in: S. Hegselmann, A. Parziale, D. Shanmugam, S. Tang, M. N. Asiedu, S. Chang, T. Hartvigsen, H. Singh (Eds.), Proceedings of the 3rd Machine Learning for Health Symposium, volume 225 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 82–100. URL: https://proceedings.mlr.press/v225/goel23a.html.

[7] N. S. Babaiha, S. G. Rao, J. Klein, B. Schultz, M. Jacobs, M. Hofmann-Apitius, Rationalism in the face of gpt hypes: Benchmarking the output of large language models against human expert-curated biomedical knowledge graphs, Artificial Intelligence in the Life Sciences 5 (2024) 100095. URL: https://www.sciencedirect.com/science/article/pii/S2667318524000023. doi:https://doi.org/10.1016/j.ailsci.2024.100095.

[8] Mistral large 2, Mistral large 2, ???? URL: https://mistral.ai/news/mistral-large-2407.

[9] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. Tsujii, S. Ananiadou, Event extraction across multiple levels of biological organization, Bioinformatics 28 (2012) i575–i581. URL: https://doi.org/10.1093/bioinformatics/bts407. doi:10.1093/bioinformatics/bts407.

[10] J.-D. Kim, Y. Wang, T. Takagi, A. Yonezawa, Overview of Genia event task in BioNLP shared task 2011, in: J. Tsujii, J.-D. Kim, S. Pyysalo (Eds.), Proceedings of BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 7–15. URL: https://aclanthology.org/W11-1802/.

[11] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, S. Ananiadou, Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013, BMC Bioinformatics 16 (2015) S2. URL: https://doi.org/10.1186/1471-2105-16-S10-S2. doi:10.1186/1471-2105-16-S10-S2.

[12] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, Biored: a rich biomedical relation extraction dataset, Briefings in Bioinformatics 23 (2022) bbac282. URL: https://doi.org/10.1093/bib/bbac282. doi:10.1093/bib/bbac282.

[13] M. Sänger, S. Garda, X. D. Wang, L. Weber-Genzel, P. Droop, B. Fuchs, A. Akbik, U. Leser, Hunflair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools, Bioinformatics 40 (2024). URL: http://dx.doi.org/10.1093/bioinformatics/btae564. doi:10.1093/bioinformatics/btae564.

[14] W. J. Wilbur, A. Rzhetsky, H. Shatkay, New directions in biomedical text annotation: definitions, guidelines and corpus construction, BMC Bioinformatics 7 (2006) 356. URL: https://doi.org/10.1186/1471-2105-7-356. doi:10.1186/1471-2105-7-356.

[15] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology (2010). URL: https://doi.org/10.1348/000711006X126600. doi:10.1348/000711006X126600.

[16] M. T. Cibulka, M. J. Strube, The conundrum of kappa and why some musculoskeletal tests appear unreliable despite high agreement: A comparison of cohen kappa and gwet ac to assess observer agreement when using nominal and ordinal data, Physical Therapy 101 (2021) pzab150. URL: https://doi.org/10.1093/ptj/pzab150. doi:10.1093/ptj/pzab150.

[17] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia Medica 22 (2012) 276–282. URL: https://www.biochemia-medica.com/en/journal/22/3/10.11613/BM.2012.031. doi:10.11613/BM.2012.031.