Understanding False Positives in Ontology-Grounded Concept Recognition: A Passage-Grounded Diagnosis

Shanshan Liu[1,2] Noriki Nishida[1] Fei Cheng[3] Narumi Tokunaga[1]
Rumana Ferdous Munne[1] Takehito Utsuro[2] Yuji Matsumoto[1]
[1]RIKEN, AIP [2]University of Tsukuba [3]Kyoto University
{shanshan.liu, noriki.nishida, narumi.tokunaga, rumanaferdous.munne}@riken.jp
feicheng@i.kyoto-u.ac.jp ; utsuro@iit.tsukuba.ac.jp ; yuji.matsumoto@riken.jp

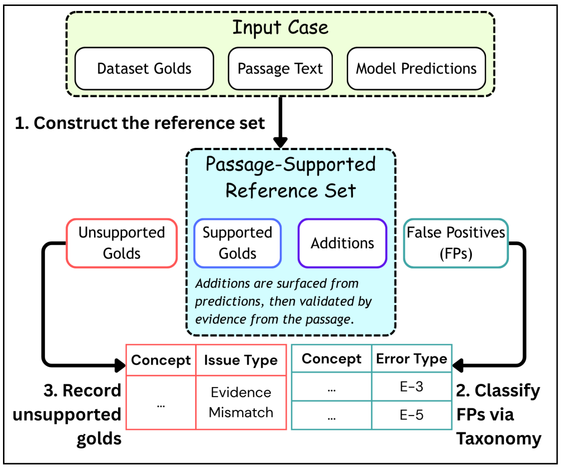
Abstract

Large language models (LLMs) support scalable auto-labeling for ontology-grounded concept recognition. However, their outputs often contain unsupported concepts and remain difficult to diagnose systematically. In this work, we introduce FP-TAX, a five-category taxonomy that explains how unsupported predictions are generated, and propose a passage-grounded diagnostic workflow that refines legacy gold annotations into a passage-supported reference set before categorizing false positives. In a case study on 20 LLM-annotated GO-BP instances, 17.3% of apparent false positives are valid additions missing from the dataset-provided gold set, while true errors are dominated by context over-generalization and inferential overreach.

1 Introduction

Biological concept recognition (BCR) maps a text passage to predefined ontology concepts, enabling scalable biomedical information extraction and downstream applications such as knowledge base construction and literature curation [1, 2, 3, 4, 5]. However, the performance of supervised BCR systems is highly contingent on the scope and quality of labeled training data. This poses a fundamental bottleneck for large ontologies (such as the Gene Ontology (GO)) – the existing labeled datasets cover only a small subset of concepts, resulting in limited generalization of models beyond the annotated subset [5, 6].

Large language models (LLMs) offer a promising alternative: the substantial presence of biomedical literature within pre-training corpora enables these models to capture domain-specific associations, catalyzing the development of scalable automated annotation pipelines to expand ontology coverage beyond manual efforts. However, recent evidence suggests that current LLM-based auto-labeling remains far from reliable: even carefully engineered pipelines achieve only 27.6% micro-F1 on GO Biological Process (GO-BP) annotation [7, 8]. This performance gap suggests that broad literature exposure alone is insufficient for high-fidelity annotation, and that the mechanisms behind these errors remain poorly understood.


Input Case
Dataset Golds
Passage Text
Model Predictions
1. Construct the reference set
Passage-Supported Reference Set
Unsupported Golds
Supported Golds
Additions
False Positives (FPs)
Additions are surfaced from predictions, then validated by evidence from the passage.
3. Record unsupported golds
2. Classify FPs via Taxonomy

Concept	Issue Type
...	Evidence Mismatch

Concept	Error Type
...	E-3
...	E-5

Figure 1 Workflow of false positive (FP) error diagnosis. We first construct a passage-supported reference set by filtering unsupported gold concepts and adding passage-supported concepts surfaced from model predictions. Predictions excluded by this reference set are treated as true FPs and categorized via FP-TAX; unsupported gold concepts are recorded as gold issues.

Enhancing automated annotation requires a granular understanding of error generation mechanisms. Do these errors stem from specific lexical triggers, or overinterpretation beyond the provided evidence? To enable systematic diagnosis, we introduce **FP-TAX**, a five-category taxonomy for false positives in ontology-grounded concept recognition, coupled with a passage-supported reference set construction procedure. FP-

TAX distinguishes between *granularity mismatch, semantic scope shift, context over-generalization, lexical triggering*, and *inferential overreach*.

Using FP-TAX, we analyze 20 LLM-annotated GO-BP instances sampled from HoIP-MLD (adapted from the CRAFT corpus) [6, 8], following the diagnostic workflow illustrated in Figure 1. The passage-supported reference set refinement reclassifies 17.3% of initially flagged false positives as passage-supported concepts missing from the dataset-provided golds, highlighting substantial gold incompleteness. Among the remaining true errors, context over-generalization (29.8%) and inferential overreach (26.9%) dominate, showing that evidence over-interpretation is the primary failure mode.

Our contributions are:

1. We introduce **FP-TAX**, a five-category taxonomy that characterizes how unsupported predictions are generated in ontology-grounded BCR.
2. We propose a passage-grounded diagnostic workflow for false positive analysis in ontology-grounded BCR, based on passage-supported reference set refinement and FP-TAX.
3. We apply the pipeline to 20 LLM-annotated GO-BP instances, quantify the false positive distribution, and provide practical design directions for improving large-scale auto-labeling.

2 Methodology

To enable the systematic characterization of erroneous predictions in biological process annotation, we introduce **FP-TAX**, a diagnostic framework consisting of five distinct error categories. This taxonomy shifts the focus from binary correctness to the underlying mechanisms of error generation.

2.1 Passage-Supported Reference Set

Before categorization, model predictions are reconciled against a passage-supported reference set. This reference set is derived by revising the dataset-provided gold annotations: specifically, we filter out gold concepts lacking textual support and incorporate **additions**—biological process concepts that are explicitly stated or unambiguously implied in the passage but omitted from the original golds.

As illustrated in the diagnostic workflow (Figure 1), these additions are surfaced from model predictions and subsequently validated by textual evidence within the passage. Consequently, a prediction is treated as a false positive (FP) only if it is excluded by this refined reference set. This protocol decouples the error analysis from the inherent incompleteness of legacy datasets, ensuring that the diagnosis remains strictly evidence-based.

2.2 FP-TAX: The Five-Tier Taxonomy

The FP-TAX categories define **how** an unsupported prediction diverges from the passage evidence, rather than merely noting the discrepancy.

E-1: Granularity Mismatch A granularity mismatch occurs when the passage supports a specific concept, but the prediction selects a term that is hierarchically related in the ontology (an ancestor or descendant node) while failing to match the precise level of abstraction evidenced in the text. These errors are strictly confined to the same ontology lineage; the model identifies the correct conceptual path but fails to calibrate the node's depth relative to the textual evidence.

E-2: Semantic Scope Shift This category applies when a prediction deviates from the semantic domain of the target concept to an unrelated class. Unlike the hierarchical depth errors in E-1, a scope shift represents a thematic misalignment where the prediction belongs to an entirely different conceptual branch. For instance, misidentifying a *learning* event as a *memory* event changes the underlying biological phenomenon being documented, thereby misrepresenting the specific functional activity characterized in the passage.

E-3: Context Over-generalization This error arises when the presence of an isolated observation or a specific biological outcome is misinterpreted as sufficient evidence for a broader, encompassing concept that is not itself stated. In these cases, the textual evidence may describe a necessary component or a resultant state of a process, but the existence of the process as a whole is not grounded in the passage. Unlike the hierarchical positioning in E-1, E-3 represents an inductive leap where partial information is erroneously treated as the occurrence of a complex, multi-stage event.

E-4: Lexical Triggering Lexical triggering refers to predictions driven by the presence of salient keywords or surface patterns rather than by the actual conceptual information. These errors reflect a reliance on stereotypical associations, where the mention of a domain-specific term—such as the name of a gene or protein (*e.g., "Shh"*) or a specific body part (*e.g., "testis"*)—biases the model toward predicting an associated concept (*e.g., "gene expression"* or *e.g., "male gonad development"*) that is not explicitly supported by the text. Here, the model's internal statistical bias regarding the keyword overrides the contextual evidence.

E-5: Inferential Overreach Inferential overreach captures predictions that necessitate unstated logical steps or external assumptions beyond the explicit description. While E-3 over-generalizes from an observed outcome to a broader conceptual class, E-5 adds entirely new explanatory logic that is absent from the text. This occurs when the model presupposes missing intermediate connections, moving from a simple descriptive statement to an unsupported causal chain.

2.3 Diagnostic Criteria

To ensure consistent classification, we distinguish the five categories via a three-step decision logic:

1. **Lineage Calibration:** If the prediction and a supported concept share the same ontological path (i.e., one is an ancestor of the other), is the error solely due to the node's depth? (If yes → **E-1**)
2. **Domain Alignment:** Does the prediction shift to a fundamentally different semantic class or an unrelated branch within the ontology? (If yes → **E-2**)
3. **Evidence Gap Analysis:** Is the lack of support due to over-extending a partial observation (**E-3**), keyword-driven bias (**E-4**), or the addition of an unsupported causal chain (**E-5**)?

3 Error Diagnosis

3.1 Experimental Setup

To investigate the structural limitations of LLM-based auto-labeling, we conducted a manual error diagnosis on the outputs generated by the pipeline from [8]. The evaluation was performed on a random sample of 20 instances drawn from the 500-instance subset of a dataset adapted from the CRAFT corpus) [6].

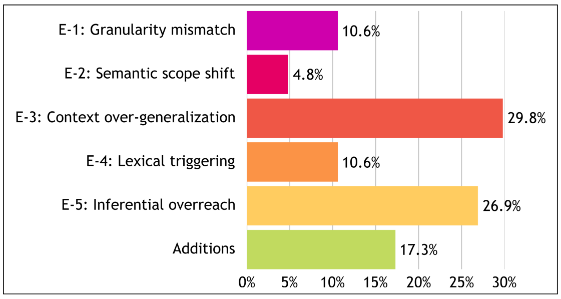


Figure 2 Distribution of initial false positives provided by an LLM-based GO-BP auto-labeling pipeline (n = 104). "Additions" are passage-supported concepts missing from dataset gold; remaining false positives are grouped into FP-TAX error types.

The necessity for this manual diagnosis stems from the inherent limitations of the legacy gold standards. As noted in prior work, the CRAFT corpus—annotated in 2010—suffers from historical biases, including overly broad annotations and missing contemporary GO refinements. Furthermore, the extreme sparsity of manual labels (covering only 2.35% of target concepts) makes them an insufficient baseline for robustly evaluating modern BCR systems. To address these bottlenecks, we employed the passage-supported reference set (as described in Section 2) to decouple pipeline failures from dataset omissions, covering 37 gold concepts and 104 initial false positive candidates in our analysis.

3.2 Results

As shown in Figure 2, 17.3% (18/104) of initially flagged false positives are reassigned as Additions—concepts that are explicitly stated or unambiguously implied by the passage but absent from the dataset-provided gold set. This indicates that a non-trivial fraction of apparent errors reflects gold incompleteness.

After removing Additions, the remaining true false positives (n = 86) exhibit a highly skewed distribution across FP-TAX categories. Context over-generalization (E-3) is the most frequent error (29.8%), where the model promotes localized observations (e.g., expression changes, phenotypic outcomes) into broader biological processes not described as such in the passage. Inferential overreach (E-5) is similarly prevalent (26.9%), reflecting the model's tendency to introduce unsupported mechanistic or regulatory structure (e.g., "positive regulation", "homeostasis", "signaling pathway") beyond what the text provides. Together, E-3 and E-5 account for 56.7% of true errors, suggesting that the dominant failure mode is over-interpretation of passage evidence rather than semantic drift.

The remaining error types occur less frequently: lexical triggering (E-4) accounts for 10.6%, typically driven by salient keywords that activate stereotyped ontology mappings; granularity mismatch (E-1) contributes 10.6%, where the model stays on the correct lineage but selects an inappropriate abstraction level; and semantic scope shift (E-2) is relatively rare (4.8%), indicating that the pipeline usually remains within the correct process domain even when incorrect.

3.3 Practical Implications

Our error analysis indicates that improving LLM-based ontology-grounded concept recognition should primarily target context over-generalization and inferential overreach.

To reduce context over-generalization, pipelines can incorporate a concept-support verification step that requires the passage to explicitly describe the predicted concept at the level of its defining semantics (e.g., a process being asserted, not merely an outcome or correlated attribute).

To mitigate inferential overreach, generation and filtering can prioritize minimal, passage-justified concepts and avoid adding extra explanatory structure—such as causal mechanisms, qualifiers, or higher-order regulatory/pathway claims—unless such relations are explicitly stated in the text.

The remaining errors occur less frequently but admit lightweight safeguards: lexical triggering may be reduced by discouraging stereotyped keyword-to-concept mappings; granularity mismatch can be addressed by ontology-depth calibration and hierarchical consistency checks; and semantic scope shift can be curtailed by restricting outputs to semantically compatible ontology branches.

Overall, these implications translate dominant error patterns into concrete design directions for more reliable large-scale concept annotation across domains.

3.4 Case studies

We present two representative cases to illustrate why systematic FP diagnosis requires both passage-supported reference refinement and a mechanism-based taxonomy (FP-TAX). Details of two cases are provided in Appendix A.

Case 1: Photoreceptor outer segment biology Out of 13 initial false positives, we identified 3 additions and 10 true errors (E-1: 2; E-3: 3; E-5: 5). The passage explicitly describes the building and maintenance of photoreceptor outer segment structure, yet the legacy gold contains only broad concepts (e.g., *"gene expression"*). Reference refinement therefore recovers passage-supported omissions such as *"photoreceptor cell outer segment organization"* as **Additions**. While absent from the gold set annotated in 2010, the concept *"photoreceptor cell outer segment organization"* was added to the Gene Ontology in 2011.

After this correction, FP-TAX shows that remaining false positives are largely driven by evidence over-interpretation, including scaling local observations to broader claims (context over-generalization; e.g., *"animal organ development"*) and introducing unsupported mechanistic explanations (inferential overreach; e.g., *"apoptotic process"* inferred from “degeneration”).

Case 2: Skin Pigmentation and Gradients Out of 6 initial false positives, we identified 2 additions and 4 true errors (E-3: 3; E-5: 1). This passage describes pigmentation gradients and pattern formation, but the dataset gold includes unsupported labels triggered by surface cues: *"visual perception"* from the phrase “visual boundary,” and *"parturition"* from “postnatal.” Reference refinement flags these gold issues and recovers omitted but passage-supported concepts (e.g., *"regulation of pigmentation"*). This case illustrates that controlling for legacy gold noise is necessary before interpreting error predictions.

Together, these cases highlight that reference set construction is essential for separating dataset issues from true model-predicted false positives, while FP-TAX provides a principled characterization of how unsupported predictions are generated.

4 Conclusion

We introduced a five-category taxonomy FP-TAX and a passage-grounded diagnostic workflow to analyze false positives in ontology-grounded concept recognition. In a case study on 20 LLM-annotated GO-BP (Gene Ontology - Biological Process) instances, 17.3% of apparent false positives are passage-supported additions missing from the gold set, while true errors are dominated by context over-generalization and inferential overreach. Our analysis highlights evidence over-interpretation as the primary failure mode by LLM-based auto-labeling pipeline and provides concrete directions for improving scalable auto-labeling.

References

[1] 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. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. **Database J. Biol. Databases Curation**, Vol. 2016, , 2016.

[2] Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. **Bioinformatics**, Vol. 37, No. 13, pp. 1884–1890, 2021.

[3] Qinyong Wang, Zhenxiang Gao, and Rong Xu. Exploring the in-context learning ability of large language model for biomedical concept linking, 2023.

[4] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, Peter N Robinson, and Christopher J Mungall. Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. **Bioinformatics**, Vol. 40, No. 3, p. btae104, 02 2024.

[5] Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. An evaluation of gpt models for phenotype concept recognition. **BMC Medical Informatics and Decision Making**, Vol. 24, No. 1, p. 30, 2024.

[6] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner Jr, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. Concept annotation in the craft corpus. **BMC bioinformatics**, Vol. 13, No. 1, p. 161, 2012.

[7] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 930–957, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Shanshan Liu, Noriki Nishida, Fei Cheng, Narumi Tokunaga, Rumana Ferdous Munne, Yuki Yamagata, Kouji Kozaki, Takehito Utsuro, and Yuji Matsumoto. Better generalizing to unseen concepts: An evaluation framework and an llm-based auto-labeled pipeline for biomedical concept recognition. In **Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, Rabat, Morocco, March 2026. Association for Computational Linguistics.

A Case Study Details

We provide the full textual evidence and model predictions for the two cases analyzed in Section 3.4 in Table 1. To assist in navigation, we have marked specific excerpts and their corresponding FP-TAX classifications.

Table 1 Detailed comparison of Passage, Dataset Golds, and LLM Predictions for Case 1 and Case 2.

Case 1	**Photoreceptor Development in $Crx^{-/-}$ Mice**
Passage	In developing photoreceptors, an extraordinary growth process occurs whereby the outer segment is generated from the nascent connecting cilium (see [38] and references therein). Peripherin/RDS and ROM-1 proteins (localized in disc rims) and the opsin proteins (localized throughout the discs) have important roles in the structural integrity of mature outer segments (see [39,29]). ROM-1-/- mice produce disorganized outer segments with large disks [40]. Crx, by virtue of being a transcription factor, presumably controls genes that are responsible for the building and perhaps maintenance of the outer segment structure, including rhodopsin and peripherin. Using northern blots [34], microarrays [10], and serial analysis of gene expression (SAGE) [35], we have defined a large number of genes that are altered in their expression level in $Crx^{-/-}$ mice. We found that rhodopsin expression was severely diminished in $Crx^{-/-}$ animals, and peripherin mRNA was reduced by approximately 30%. Transgenic mice with variable levels of expression of wild type rhodopsin exhibit rod degeneration [41], indicating the importance of the level of rhodopsin expression. In addition, the timing of rhodopsin expression may be very important, as indicated by studies in Drosophila.
Gold	*"biological regulation"*, *"gene expression"*
Prediction	**[Additions]**: *"photoreceptor cell outer segment organization"*, *"regulation of gene expression"*, *"developmental process"* **[E-1]**: *"photoreceptor cell development"*, *"photoreceptor cell morphogenesis"* **[E-3]**: *"animal organ development"*, *"anatomical structure homeostasis"*, *"biological phase"* **[E-5]**: *"cilium assembly"*, *"execution phase of apoptosis"*, *"apoptotic process"*
Case 2	**Skin Pigmentation Gradients in at/at Mice**
Passage	The visual boundary between dorsal and ventral skin in at/at mice is reminiscent of other systems in which adjacent compartments enforce a binary choice between alternative patterns of gene expression and cell fate (reviewed in Dahmann and Basler 1999). However, Agouti mRNA in both embryonic and postnatal skin is distributed along a gradient whose dorsal boundary is indistinct and overlaps with two additional gradients recognized by their effects on hair length and histochemical staining for melanocytes. The three gradients are close but not congruent, and it is their proximity that gives rise to the superficial distinction between dorsal and ventral skin of at/at mice. Indeed, slight differences between the regions of transition for pigment-type switching and pigment content give rise to a subtle yellow stripe along the flank (see Figures 1, 2, and 9A). Levels of Agouti mRNA remain high throughout the entire ventrum, but hair pigment content is reduced, giving rise to a cream-colored region in the ventrum that, depending on age and genetic backgrounds, may appear more or less distinct from the yellow flank stripe.
Gold	*"gene expression"*. **[Unsupported Golds]**: *"visual perception"* (E-4), *"parturition"* (E-3).
Prediction	**[Additions]**: *"regulation of pigmentation"*, *"pattern specification process"* **[E-3]**: *"post-embryonic development"*, *"pigmentation"*, *"developmental process"* **[E-5]**: *"polyphenic determination, influence by genetic factors"*