Ground-truth Construction and Evaluation of LLM Contribution to Life Sciences Tool Annotation

Ulysse Le Clanche[1,*], Melvin Selim Atay[2], Elise Bannier[2,3], Anaïs Baudot[4,5], Lea Bellenger[6], Alexandrina Bodrug[6], Samuel Chaffron[7], Eric Charpentier[6], Erwan Corre[8], Clémence Frioux[9], Aurélie Lardenois[10], Frédéric Lemoine[11], Camille Maumet[2], Cyril Noël[12], Perrine Paul-Gilloteaux[13], Paul Simion[14], Morgane Térézol[4], Olivier Dameron[1,*] and Alban Gaignard[6,*]

[1]*Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France*
[2]*Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074, EMPENN — ERL U 1228, F-35000 Rennes, France*
[3]*CHU Rennes, Radiology Department, Rennes France*
[4]*Aix Marseille Université, INSERM, MMG, Marseille, France*
[5]*CNRS, Marseille, France*
[6]*Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France*
[7]*Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*
[8]*ABiMS-IFB, Station Biologique de Roscoff, CNRS/Sorbonne Université, Roscoff, France*
[9]*Inria, Univ. Bordeaux, INRAE, 33400, Talence, France*
[10]*Institut National de Santé et de Recherche Médicale, U1085-Irset, Université de Rennes 1, F-35042 Rennes, France*
[11]*Institut Pasteur, Université Paris Cité, Bioinformatics of Biostatistics Hub, F-75015 Paris, France*
[12]*SeBiMER Service de Bioinformatique de l'Ifremer, Ifremer, IRSI, Plouzané, France*
[13]*Nantes Université, CHU Nantes, CNRS, Inserm, BioCore, US16, SFR Bonamy, Nantes, France*
[14]*EcoBio - Ecosystems, Biodiversity, Evolution, Université de Rennes 1 35042 Rennes, France*

Abstract

Metadata are explicit representations of the meaning of data or of what aspects of the data are important. Their quality directly impacts data retrieval and exploitation, so curation across the whole data life cycle is critical. Curation complexity is such that human experts need assistance from automatic tools, but the design of such tools is challenging. In this work, we compare the respective capabilities of domain experts and of an Large Language Model (LLM) agent to generate useful domain-specific annotations for classifying, indexing, and searching scientific resources. We ground this question in the life sciences domain with the bio.tools software registry and the companion EDAM reference ontology. Our contribution is twofold: i) we propose a ground-truth dataset of semantically annotated life sciences software, created and validated by a set-group of experts, and ii) we study the relative contribution of an LLM agent to this ground-truth, compared to a group of domain experts. Our results show the complementarity between domain experts and LLMs: most annotations from the ground-truth were provided by the experts, but the LLM also contributed annotations that were overlooked by the experts. Some relevant annotations provided by either experts or LLM do not correspond to any class in the reference ontology, highlighting the need for curation at the level of both metadata and ontologies.

Keywords

Semantic annotations, LLM agent, annotation curation, ground-truth, life sciences

CEUR-WS.org/Vol-4246/11ms4kgoe-1.pdf

LLM4KGOE'26: Workshop on LLM-driven Knowledge Graph and Ontology Engineering, Co-located with ESWC 2026
*Corresponding author.
ulysse.le-clanche@irisa.fr (U. Le Clanche); olivier.dameron@irisa.fr (O. Dameron); alban.gaignard@univ-nantes.fr (A. Gaignard)
0009-0004-3536-985X (U. Le Clanche); 0000-0003-4985-2362 (M. Atay); 0000-0002-8942-7486 (E. Bannier); 0000-0003-0885-7933 (A. Baudot); 0000-0002-5332-7870 (L. Bellenger); 0009-0006-3873-102X (A. Bodrug); 000-0001-5903-617X (S. Chaffron); 0000-0002-8571-7603 (E. Charpentier); 0000-0001-6354-2278 (E. Corre); 0000-0003-2114-0697 (C. Frioux); 0000-0003-1339-104X (A. Lardenois); 0000-0001-9576-4449 (F. Lemoine); 0000-0002-6290-553X (C. Maumet); 0000-0002-7139-4073 (C. Noël); 0000-0002-4822-165X (P. Paul-Gilloteaux); 0000-0002-6667-3297 (P. Simion); 0000-0002-4090-2573 (M. Térézol); 0000-0001-8959-7189 (O. Dameron); 0000-0002-3597-8557 (A. Gaignard)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-08-26

1. Introduction

Data annotations, also called metadata, are explicit representations of the meaning of data or of what aspects of the data are important. They are a proxy to overcome both data inner complexity and relations between data elements [1]. The quality of annotations directly impacts data retrieval and exploitation [2], so curation across the whole data life cycle is critical [3]. Curation is challenging because it draws on information and background knowledge that are not always present in organized databases, but scattered across textual documentation, websites, or even multiple experts' minds. This complexity makes it primarily a human-driven task, difficult to fully automate, although human experts may benefit from computational assistance [4]. In this work, we compare the capabilities of domain experts and an LLM agent in generating useful domain-specific annotations for better classifying, indexing, and searching scientific resourcess, focusing on life sciences software. Bio.tools [5] is a widely used catalogue describing over 30,000 life sciences software tools. It aggregates rich metadata such as versions, documentation, credit, and licenses, improving tool findability, accessibility, and reuse. It relies on the EDAM ontology [6, 7] as its semantic backbone. EDAM organizes life sciences knowledge into four sub-ontologies: *EDAM Topic* (scientific domains), *EDAM Operation* (what tools do), *EDAM Data* (nature of input/output data), and *EDAM Format* (data representation). While registering a new software in bio.tools is relatively fast and easy, providing complete and precise EDAM annotations requires substantial expertise and effort. Our previous work [8] showed that EDAM annotations in bio.tools exhibit a high variability in both quantity and informativeness. Without a ground-truth, evaluating annotation quality and prioritizing curation efforts remains difficult. **The contribution of this work is twofolds: i) we propose a ground-truth dataset of semantically annotated life sciences software, created and validated by a set-group of experts, and ii) we study the relative contribution of an LLM agent to this ground-truth, compared to a group of domain experts.** The remainder of this article is organized as follows. Section 2 presents the state of the art. Section 3 introduces the ground-truth creation pipeline. Section 4 reports the results, including the dataset and analysis of LLM contributions. Section 5 discusses these results in light of the state of the art, and Section 6 concludes with findings and future directions.

2. State of the art

A growing number of studies explores LLMs for metadata extraction and knowledge graph construction. Turner *et al.* [9] demonstrate that LLMs can automatically extract structured metadata from neuroimaging publications, enabling semi-automated annotation pipelines. However, their results highlight that expert validation remains necessary to ensure semantic correctness and reliability. In ontology engineering, Norouzi *et al.* [10] show that LLMs can generate candidate entities and relations for ontology population, but emphasize the dependence on curated ground-truth datasets and human supervision. Similarly, Smit *et al.* [11] achieve strong performance (F1-score 0.94) for bioassay annotation using a fine-tuned NER model trained on 800 manually curated assays, but at the cost of substantial labeled data requirements. More recently, Riquelme-García *et al.* [12] evaluate base and fine-tuned LLMs for ontology-based biological annotation, showing large performance variability (F1-score 19%–77%) depending on the ontology and annotation type. These works confirm the promise of LLMs for semantic annotation, but also consistently highlight three limitations: dependence on labeled data, variability across tasks, and the need for expert validation.

3. Ground-truth creation

In this section, we propose a workflow leveraging expert knowledge to build a reference dataset of life sciences software annotations describing functionalities, input/output types, and contexts of use. Figure 1 outlines the main steps. After individual tool annotation, experts reach a consensus for a given domain. In parallel, an LLM agent annotates the tools. Experts assess the relevance of these annotations

and produce a new consensus. Finally, annotations are manually semantically aligned with the EDAM ontology by experts. The resulting dataset is a consensus-based set of EDAM annotations incorporating relevant LLM contributions.

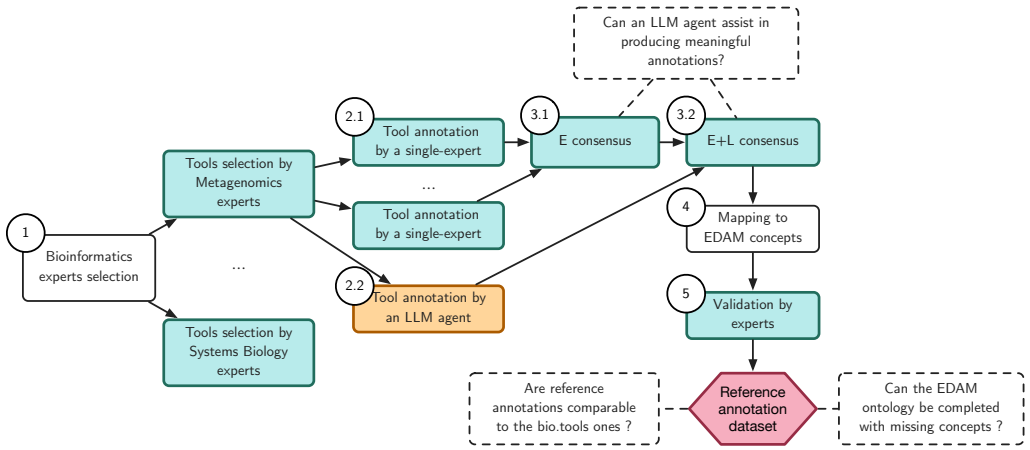

Figure 1: Ground-truth dataset building workflow, combining LLM-based and expert contributions. Green boxes represent expert activities, orange boxes LLM agent contributions, and white boxes manual steps. Dotted boxes indicate research questions. In Step 3.1, experts meet to verify and establish a consensus ("E consensus"). In Step 3.2, experts evaluate the relevance of LLM-generated terms, leading to an "E+L consensus" (expert and LLM consensus).

Step 1: Scope definition. We identified 7 subdomains covering various aspects of life sciences, with at least two experts per subdomain. Experts are scientists selected for their familiarity with the tools and practices of their specific field. The aim is not exhaustive domain coverage, but a focused set of well-established, practice-oriented subdomains. In total, 16 experts participated: 2 in Genetic Variant, 2 in Metagenomics, 3 in Phylogeny, 2 in Single Cell, 3 in Systems Biology, and 4 in Bio-imaging (Microscopy and Neuroimaging). Each expert is associated with a single domain of expertise. Experts were asked to individually identify five reference tools from their field of daily practice.

Step 2.1 Individual annotation by experts. For each tool, we asked experts to provide free text annotations describing the associated scientific discipline(s), the tool's functionality, the nature of input and output data, and the data formats used. Experts were instructed not to consult domain ontologies in order to allow the identification of potentially missing concepts, and not to rely on LLM-based agents for annotation.

Step 2.2 LLM-based annotation. In Step 2.2, we asked an LLM agent to perform the same annotation task as the experts, following identical guidelines to ensure a fair and consistent comparison between human and model-generated annotations. We used DeepSeek v3.1, a free and open-source model with "Web Search" and "Deep Think" capabilities, where Web Search retrieves up-to-date web information and Deep Think enables more comprehensive reasoning by allocating additional computation time. We used the prompting strategy described on GitHub[1], where both the prompt design and illustrative examples are provided. For each annotation task, a separate DeepSeek chat was created, and results were not immediately shared with experts to avoid bias.

Step 3: Consensus. The consensus process was based on discussion. For each domain, experts first held a meeting to reach consensus on annotations (Step 3.1 *E consensus*), followed by a second meeting to evaluate LLM-generated terms (Step 3.2 *E+L consensus*). In Step 3.2, new annotations could emerge from LLM suggestions. When adopted, these were counted as mixed contributions. Experts

[1] https://github.com/ulysseLeclanche/EDAM_ground_truth/blob/main/Supplementary_material.pdf

were aware of which annotations came from the LLM. The resulting datasets enable analysis of the LLM's contribution to ground-truth annotation.

Step 4: Mapping. Free-text annotations were manually mapped to EDAM concepts using the EDAM browser tool[2], relying on its search engine for labels, synonyms, and definitions. We used EDAM version 1.25, a stable release. Each mapping was classified as exact match, synonym, broader, or narrower concept. Despite this process, some annotations could not be mapped.

Step 5: Validation. Experts then validated the mapped EDAM terms with a yes/no decision. When rejected, they provided justification and could propose alternative EDAM terms to improve the ontology.

4. Results

4.1. A ground-truth of annotated life sciences software dataset

The final ground-truth dataset includes 532 unique free-text annotations built from a consensus between domain experts and a large language model (DeepSeek). As reported in Table 1, the study covers seven scientific domains (Metagenomic, Neuroimaging, Phylogeny, Single Cell, Systems Biology, Genetic Variant, and Microscopy) and a total of 44 tools. The annotations are distributed across four types: 184 *EDAM Data*, 146 *EDAM Operation*, 105 *EDAM Topic*, and 97 *EDAM Format*. Experts contributed the majority of annotations to the final consensus, with an 84.8% retention rate and 15.2% rejection rate, while DeepSeek achieved a 41.3% retention rate and 58.7% rejection rate across all domains. LLM suggestions led experts to revise annotations in 7.6% of cases, creating new "mixed" contributions.

Table 1
Distribution of free-text annotations (Topic, Operation, Data, and Format) in the ground-truth across seven life sciences domains. For each domain, the number of free-text annotations is reported without accounting for redundancies across domains. In the "Total" row, redundancies in free-text annotations across domains are included, as well as redundancies in tools.

Domain	Nb. Tools	Topic	Operation	Data	Format
Metagenomic	8	15	25	24	10
Neuroimaging	5	10	9	17	25
Phylogeny	3	6	8	9	5
Single Cell	3	4	13	20	11
Systems biology	11	48	61	76	44
Genetic variant	10	14	18	28	19
Microscopy	4	8	12	17	14
Total (unique values)	**44**	**105**	**146**	**184**	**97**

4.2. Do LLMs miss or predict annotations differently from experts?

The LLM shows a global recall of 0.46, meaning it misses about half of the annotations identified in the expert/LLM consensus, compared to a much higher expert recall of 0.91. Overall, recall is 0.46 ± 0.33 and varies strongly across domains from 0.28 ± 0.34 for Neuroimaging to 0.62 ± 0.25 for Systems Biology. It also varies between tools within the same domain, e.g. in Phylogeny from 0.145 (FastMe) to 0.443 (PhyML). The LLM achieves a moderate global precision of 0.43, indicating that a substantial proportion of its predictions are not retained in the final consensus, while expert precision is higher (0.87). The lowest LLM precision is observed in Microscopy (0.30 ± 0.29), compared to 0.87 ± 0.28 for experts. However, the LLM also proposes additional relevant annotations not initially suggested by experts, including both specific terms (e.g. *"Voxel-wise regression analysis"* for AFNI in Neuroimaging) and broader concepts (e.g. *"Molecular evolution"* for PhyML in Phylogeny). During discussions, LLM

[2]https://github.com/edamontology/edam-browser

suggestions may also trigger new annotations not initially proposed by either experts or the LLM. Precision also varies across domains, with the highest LLM precision in Systems Biology (0.72 ± 0.33) and values often below or equal to 0.5 elsewhere. The high variance across tools within domains further explains the overall dispersion, e.g. in Neuroimaging from 0.20 (fMRIPrep) to 0.46 (SPM). *Experts play a crucial role for determining the correct annotations, but assistance helps assuring exhaustiveness. DeepSeek provides valuable annotations that experts overlook, even though its overall accuracy remains moderate.*

4.3. Are LLMs better annotators than Experts (and conversely)?

4.3.1. Global F1 score analysis

The relatively large standard deviations observed for LLM annotations (F1-score 0.417 ± 0.305) reflect strong variability across tools and annotation types within the same domain. Experts achieve a significantly higher F1-score (0.893 ± 0.183) than DeepSeekV3.1. Overall, LLMs are less accurate than experts in predicting annotations, with lower precision (0.434 vs. 0.901), recall (0.458 vs. 0.906), and F1-score (0.417 vs. 0.893). They also do not necessarily predict the same annotations: DeepSeek tends to generate more annotations, but these are less accurate and often miss annotations identified by experts. While experts consistently outperform the LLM, both do not produce the same annotation sets, with DeepSeek providing a larger but less accurate set that fails to capture part of the expert-proposed annotations.

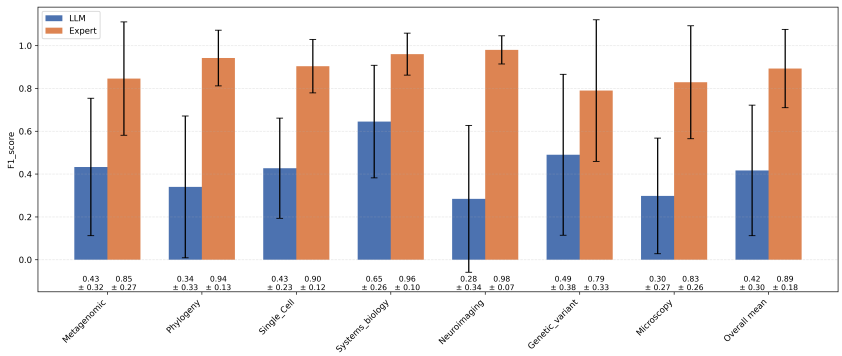

Figure 2: F1 score comparison between LLM and experts. F1-score values correspond to the mean F1-score computed over all tool annotations within each domain, along with their standard deviation. The figure shows that the highest LLM F1-score is observed in Systems Biology (0.65 ± 0.26), where it is half in Phylogeny (0.34 ± 0.33).

The domain-level analysis reveals strong variations in LLM performance across scientific fields (Figure 2). The best results for the LLM are observed in Systems Biology (mean F1 = 0.65), where both precision (0.72) and recall (0.62) are relatively high. It also achieves comparatively better performance in Genetic Variant (mean F1 = 0.49), mainly driven by higher recall (0.62). In contrast, performance is much lower in Neuroimaging (mean F1 = 0.28), Phylogeny (mean F1 = 0.34), and Microscopy (mean F1 = 0.30). Metagenomic (mean F1 = 0.43) and Single Cell (mean F1 = 0.43) show intermediate performance, though still with considerable variability, as reflected by relatively large standard deviations across domains. Expert performance, however, remains high across all domains, with consistently strong stronger F1-scores and lower variability (± 0.18) than LLM (± 0.30). *Experts contribute more to the consensus annotations, they perform better and are more consistent than the LLM. However in certain domains, DeepSeek still provides some additional annotations retained by experts. The contribution of LLM to annotation clearly varies depending on the domain, with better performance in the field of systems biology and genetic variants, and less effective in fields such as Neuroimaging, Phylogeny, or Microscopy.*

4.3.2. Do LLM performances depend on the type of annotation?

The analysis by annotation type shows that LLM performance varies depending on the category of annotation (Figure 3). Two main groups emerge. The first includes *formats*, with both recall and precision above 0.6, and input data annotations, which also reach values above 0.5 for both metrics. In

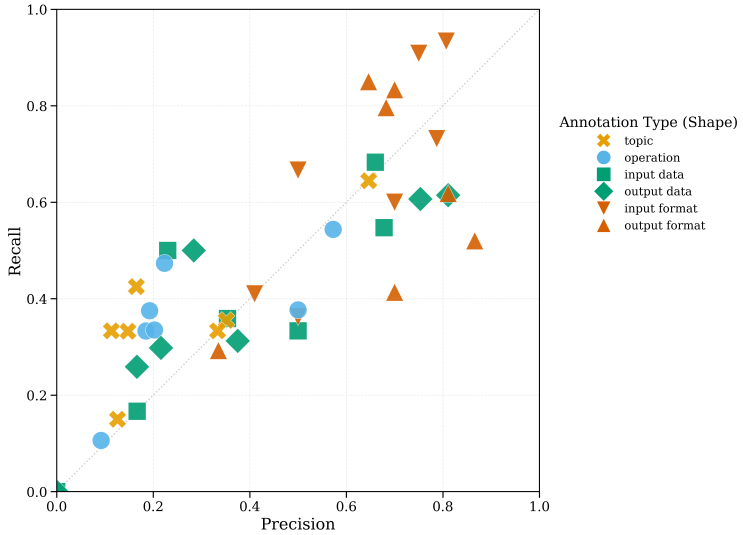

Figure 3: Precision and Recall for LLM contribution by annotation type. Each dot represent the aggregation of the annotations of a particular type of the tools of a domain. For each annotation type, we computed the average precision and recall across all tools generated by the LLM and validated through expert–LLM consensus.

contrast, most operation and topic annotations show lower performance, with both precision and recall generally below 0.4. This suggests that a general LLM such as DeepSeek is less effective at proposing relevant *Operations* or *Topics*, often generating general terms that are rejected by experts. It performs better on more concrete annotations such as input and output formats.

The LLM's contribution varies by annotation type, as it performs better on format identification, but struggles with fine-grained or specific annotations especially on topics and operations annotation type.

4.4. Which annotations identified by the LLM and experts were missing in EDAM?

Table 2
Mapping of free-text annotations with EDAM ontology classes Missing EDAM classes are identified as "No candidate". Other free-text annotations can be either unvalidated or assigned to an EDAM class through *exact*, *broader*, or *narrower* mappings respectively when the annotation is equivalent, more precise/general, and more general/precise than the EDAM class

Domain	No candidate	Unvalidated	Exact	Broader	Narrower
Genetic variant	0	3	25	48	6
Metagenomic	0	19	30	20	4
Microscopy	0	10	10	31	0
Neuroimaging	5	8	10	25	0
Phylogeny	0	2	10	17	2
Single Cell	3	6	16	22	2
Systems biology	24	59	62	88	1
Total	32	107	163	251	15

Out of the 44 ground-truth tools proposed by experts, 6 (2 in Genetic Variant, 2 in Microscopy, 1 in Neuroimaging, and 1 in Systems Biology) are not present in the bio.tools database. The 28 unique EDAM annotations associated with these tools and validated by experts were therefore removed to enable comparison with bio.tools. Among all free-text annotations proposed by experts and the LLM, most were validated, resulting in 429 non-unique EDAM annotations (Table 2). Of these, 56% (240 annotations) correspond to EDAM classes broader than those proposed by experts, while only 26 correspond to more specific classes. This indicates that the EDAM ontology lacks sufficiently specific classes to describe life sciences tools, particularly in domains such as Neuroimaging, Single Cell, and Systems Biology. This

limitation is further reflected in the 32 free-text annotations with no matching EDAM terms, including examples such as *"brik/head"* (format, Neuroimaging), *"ATAC-seq"* (data, Single Cell), or *"similarity matrices creation"* (operation, Systems Biology). Some non-validated annotations could be considered as candidates for extending EDAM. An annotation may be rejected because it is too general to be informative, or because, its definition or position in the EDAM hierarchy is not appropriate.

Most of the experts' free-text annotations matched an EDAM class, although EDAM classes were often more general. This shows the need to expand the EDAM ontology to describe tools more precisely. Currently, EDAM does not cover all fields equally. Some areas, especially Neuroimaging, Single-Cell and Systems Biology, lack suitable EDAM terms for their tools.

4.5. Added value for the bio.tools repository: which annotations identified by the LLM and experts were missing in bio.tools?

Of the 38 tools shared between the ground-truth and bio.tools, Table 3 shows a 29% agreement between annotations. Additionally, 38% are specific to the ground-truth, suggesting they could enrich bio.tools, while the remaining 32% are specific to bio.tools and may correspond either to relevant annotations missed by experts or to less reliable ones possibly originating from automated methods.

Table 3
Comparison of EDAM annotations validated by experts (GT: ground-truth) and annotations from tools in bio.tools.

	GT only	GT ∩ Bio.tools	Bio.tools only
Stated annotations	102 (38%)	78 (29%)	87 (32%)
Stated and inferred annotations	112 (26%)	204 (47%)	122 (28%)

Stated annotations correspond to expert-validated ground-truth annotations and those available in the bio.tools database. However, these comparisons do not take into account the semantic precision of the annotations: for instance, an annotation appearing specific to the ground-truth (e.g. *Sequence assembly*, `edam:0310`) may actually be a superclass of a more specific term in bio.tools (e.g. *De-novo assembly*, `edam:0524`), highlighting the importance of considering the EDAM hierarchy to avoid underestimating agreement. Inferred annotations are derived from stated annotations using the EDAM ontology. When inferred annotations are included, agreement increases from 29% to 47%. Consequently, 26% of annotations could contribute as meaningful additions to bio.tools, while the remaining 28% require further assessment of their relevance. *Overall, two thirds of the inferred annotations from the ground-truth were already present in bio.tools, and one third are relevant annotations that should be added. Conversely, two thirds of the annotations from bio.tools were confirmed by the experts. The relevance of the remaining annotations from bio.tools cannot be determined at this point.*

5. Comparison to state of the art

Like Turner *et al.* [9], we used a general-purpose LLM with web search capabilities to extract annotations across multiple domains, although we manually mapped EDAM classes to free-text annotations. Our approach aligns with Norouzi *et al.* [10]'s recommendations, which provide a curated ground-truth dataset and an expert-driven methodology. In contrast to approaches relying on large manually annotated corpora for training (e.g., Smit *et al.* [11]), our work leverages off-the-shelf LLMs to construct a ground truth, avoiding the need for prior labeled datasets which do not exist in our domain. As in Riquelme-García *et al.* [12], we observe that performance varies across annotation types, particularly for broad or ambiguous classes. Our results are consistent with these findings and further confirm that LLM-based annotation systems benefit from expert validation to ensure reliability.

6. Discussion and conclusion

In this work, 16 life sciences experts manually annotated 44 tools through an iterative process involving multiple consensus phases and LLM-generated proposals. This resulted in a ground-truth dataset of 532 unique annotations: 184 *EDAM Data*, 146 *EDAM Operations*, 105 *EDAM Topics*, and 97 *EDAM Formats*. Experts contributed most of the retained annotations, with a mean retention rate of 84.8%, and achieved high overall recall (0.907 ± 0.231) and precision (0.885 ± 0.247), consistently across domains.

In comparison, the LLM showed lower and more variable performance, with moderate precision (0.480 ± 0.376) and recall (0.458 ± 0.332), and a rejection rate of 58.7%. While some suggestions lacked accuracy, others were relevant and complementary to expert annotations. These results indicate that LLMs cannot replace experts but can support the annotation process.

Performance also varied across domains, with better results in Systems Biology and Genetic Variants, and lower performance in Neuroimaging, Phylogeny, and Microscopy. This heterogeneity is difficult to interpret given limited transparency on LLM training data, although all tools were available prior to training, suggesting that performance differences are not due to unseen data. In contrast, expert performance remained strong across all fields. There is potential bias, as the experts decide, during the consensus phases, which annotation to retain or reject, and a possible voting effect depending on group composition, but it reflects the situation that users and particularly domain experts are ultimately the ones who determine adequacy.

LLM performance also depends on annotation type. It performs well on format annotations, which are standardized and easily extracted, but struggles with more specific and fine-grained Topic and Operation annotations, often producing general terms. This highlights its relative strength on less abstract information and its limitations on specialized knowledge. Conversely, experts tend to overlook annotation Formats and Data in favor of more precise Topics and Operations.

The ground-truth dataset can enrich bio.tools by adding 102 annotations for 38 existing tools. Additionally, 6 tools and their 28 annotations are not yet present in the database despite regular expert use. The study also identifies gaps in the EDAM ontology, with 32 missing classes in domains such as Neuroimaging, Single Cell, and Systems Biology. Among 107 non-validated annotations, some may serve as candidates for extending EDAM.

Overall, these results highlight the complementarity between experts and LLMs: experts ensure accuracy and consistency, while LLMs help identify additional annotations and reveal possible ontology gaps.

Although this study provides insights into the use of LLMs as complementary annotators, it is limited to a single model and a restricted set of domains and tools. Results may differ with other models or broader datasets. A bio-imaging extension of EDAM is currently under development, emphasizing the need to continuously update such ground-truth resources. Future work includes evaluating Retrieval-Augmented Generation [13] (RAG)-based annotation methods and exploring *LLM-as-a-judge* [14] approaches. To reduce the cost of human-based annotation, we aimed at using this ground-truth to better measure annotation quality and identify guidelines for annotation best practices.

Acknowledgments

This work was supported in part by the National Research Agency under the France 2030 program, with reference to ANR-22-PESN-0007 (ShareFAIR project).

Dataset and source code availability

The ground-truth dataset, prompts, and analysis source code are available on GitHub[3]. The EDAMannot[4] toolbox was used to extract annotations from bio.tools and to infer additional annotations.

[3] https://github.com/ulysseLeclanche/EDAM_ground_truth
[4] https://github.com/ulysseLeclanche/EDAMannot

References

[1] C. Bizer, T. Heath, T. Berners-Lee, Linked data - the story so far, Int. J. Semantic Web Inf. Syst. 5 (2009) 1–22.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 160018.

[3] C. Goble, R. Stevens, D. Hull, K. Wolstencroft, R. Lopez, Data curation + process curation=data integration + science, Briefings in bioinformatics 9 (2008) 506–517.

[4] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, L. Hunter, Manual curation is not sufficient for annotation of genomic databases, Bioinformatics (Oxford, England) 23 (2007) i41–i48.

[5] J. Ison, H. Ienasescu, P. Chmura, E. Rydza, H. Ménager, M. Kalaš, V. Schwämmle, B. Grüning, N. Beard, R. Lopez, S. Duvaud, H. Stockinger, B. Persson, R. S. Vařeková, T. Raček, J. Vondrášek, H. Peterson, A. Salumets, I. Jonassen, R. Hooft, T. Nyrönen, A. Valencia, S. Capella, J. Gelpí, F. Zambelli, B. Savakis, B. Leskošek, K. Rapacki, C. Blanchet, R. Jimenez, A. Oliveira, G. Vriend, O. Collin, J. van Helden, P. Løngreen, S. Brunak, The bio.tools registry of software tools and data resources for the life sciences, Genome Biology 20 (2019) 164. URL: https://doi.org/10.1186/s13059-019-1772-6. doi:10.1186/s13059-019-1772-6.

[6] J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, P. Rice, EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats, Bioinformatics 29 (2013) 1325–1332. doi:10.1093/bioinformatics/btt113.

[7] M. Black, L. Lamothe, H. Eldakroury, M. Kierkegaard, A. Priya, A. Machinda, U. S. Khanduja, D. Patoliya, R. Rathi, T. P. C. Nico, G. Umutesi, C. Blankenburg, A. Op, P. Chieke, O. Babatunde, S. Laurie, S. Neumann, V. Schwammle, I. Kuzmin, C. Hunter, J. Karr, J. Ison, A. Gaignard, B. Brancotte, H. Menager, M. Kalas, Edam: The bioscientific data analysis ontology (update 2021), F1000Research 11 (2021). doi:10.7490/f1000research.1118900.1.

[8] U. Le Clanche, S. Cohen-Boulakia, Y. L. Cunff, O. Dameron, A. Gaignard, Assessing bioinformatics software annotations: bio.tools case-study, in: Proceedings JOBIM, Bordeaux & Online, France, 2025, pp. 1–7. URL: https://inria.hal.science/hal-05096849.

[9] M. D. Turner, A. Appaji, N. Ar Rakib, P. Golnari, A. K. Rajasekar, A. R. KV, S. S. Sahoo, Y. Wang, L. Wang, J. A. Turner, Large language models can extract metadata for annotation of human neuroimaging publications, Frontiers in Neuroinformatics 19 (2025) 1609077.

[10] S. S. Norouzi, A. Barua, A. Christou, N. Gautam, A. Eells, P. Hitzler, C. Shimizu, Ontology population using llms, 2024. URL: https://arxiv.org/abs/2411.01612. arXiv:2411.01612.

[11] I. Smit, M. Adasme, E. Manners, S. Corbett, N. Bosc, H.-M.-A. Do, A. Leach, N. O'Boyle, B. Zdrazil, Integrating artificial intelligence and manual curation to enhance bioassay annotations in chembl, Journal of Cheminformatics 18 (2026). doi:10.1186/s13321-026-01165-x.

[12] A. Riquelme-García, J. Mulero-Hernández, J. T. Fernández-Breis, Annotation of biological samples data to standard ontologies with support from large language models, Computational and Structural Biotechnology Journal 27 (2025) 2155–2167. URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037025001837. doi:10.1016/j.csbj.2025.05.020.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[14] L. Dietz, O. Zendel, P. Bailey, C. L. A. Clarke, E. Cotterill, J. Dalton, F. Hasibi, M. Sanderson, N. Craswell, Principles and guidelines for the use of llm judges, in: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), ICTIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 218–229. URL: https://doi.org/10.1145/3731120.3744588. doi:10.1145/3731120.3744588.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.5 (OpenAI) in order to: improve the clarity, and conciseness of the language. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.