Efficient and Reliable Estimation of Named Entity Linking Quality: A Case Study on GutBrainIE

Marco Martinelli[1,*], Stefano Marchesin[1] and Gianmaria Silvello[1]

[1]Department of Information Engineering, University of Padua, Italy

Abstract

Named Entity Linking (NEL) is a core component of biomedical Information Extraction (IE) pipelines, yet assessing its quality at scale is challenging due to the high cost of expert annotations and the large size of corpora. In this paper, we present a sampling-based framework to estimate the NEL accuracy of large-scale IE corpora under statistical guarantees and constrained annotation budgets. We frame Named Entity Linking (NEL) accuracy estimation as a constrained optimization problem, where the objective is to minimize expected annotation cost subject to a target Margin of Error (MoE) for the corpus-level accuracy estimate. Building on recent works on knowledge graph accuracy estimation, we adapt Stratified Two-Stage Cluster Sampling (STWCS) to the NEL setting, defining label-based strata and global surface-form clusters in a way that is independent of NEL annotations. Applied to 11,184 NEL annotations in GutBrainIE – a new biomedical corpus openly released in fall 2025 – our framework reaches a MoE ≤ 0.05 by manually annotating only 2,749 triples (24.6%), leading to an overall accuracy estimate of 0.915 ± 0.0473. A time-based cost model and simulations against a Simple Random Sampling (SRS) baseline show that our design reduces expert annotation time by about 29% at fixed sample size. The framework is generic and can be applied to other NEL benchmarks and IE pipelines that require scalable and statistically robust accuracy assessment.

Keywords

Named Entity Linking, Quality Estimation, Accuracy Estimation, Information Extraction, Gut-Brain Axis

1. Introduction

NEL is a key component in Information Extraction (IE) pipelines, especially in biomedical domains where downstream tasks such as Relation Extraction (RE) and Knowledge Base Construction (KBC) strictly depend on the correctness of entity-concept mappings [1, 2, 3]. These tasks play a crucial role in building reliable automatic IE systems that can support clinicians and researchers in processing large volumes of unstructured text, for instance when conducting biomedical literature reviews or extracting structured information from triage records and other clinical documents [4, 5, 6]. However, assessing the quality of NEL annotations at scale is challenging for two reasons: expert annotations are expensive, and manually revising all linked mentions one by one is often infeasible [7, 8].

In this work, we tackle the challenge of estimating the accuracy of NEL in large-scale IE corpora, with a focus on scalability and cost-efficiency. To overcome this, we introduce an iterative, sampling-based framework that selectively draws triples for manual verification and leverages those verifications to estimate corpus-level accuracy with formal statistical guarantees. The core idea is to repeatedly sample a small, representative subset of triples (mention-concept linkages in the present context), obtain manual judgments on them, and use these results to refine the overall accuracy estimate. This process continues until the estimate falls within a predefined margin of error. The proposed estimation framework rests on two foundational elements: establishing a statistical Confidence Interval (CI) and selecting an appropriate sampling strategy to meet that requirement efficiently. The CI specifies the

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

martinell2@dei.unipd.it (M. Martinelli); stefano.marchesin@unipd.it (S. Marchesin); gianmaria.silvello@unipd.it (G. Silvello)

https://www.dei.unipd.it/~martinell2/ (M. Martinelli); https://www.dei.unipd.it/~marches1/ (S. Marchesin); https://www.dei.unipd.it/~silvello/ (G. Silvello)

0009-0001-1596-8642 (M. Martinelli); 0000-0003-0362-5893 (S. Marchesin); 0000-0003-4970-4554 (G. Silvello)

acceptable level of uncertainty in our accuracy estimate, while the sampling strategy determines how triples are selected for manual verification to achieve this confidence level at minimal cost.

We validate this framework through a case study on the GutBrainIE collection [9, 10], a large-scale corpus on the gut-brain axis [11, 12], featuring over 11'000 concept-level linkages.

Our main contributions are:

- We formalize NEL accuracy estimation as a constrained optimization problem, where the goal is to minimize annotation cost under a target CI constraint.
- We adapt Stratified Two-stage Weighted Cluster Sampling (STWCS), a State of the Art (SOTA) sampling strategy developed for Knowledge Graph (KG) accuracy estimation [13], to the NEL task.
- We evaluate the framework on GutBrainIE [14], showing that annotating less than 25% of all triples is enough to obtain a corpus-level accuracy estimate with MoE ≤ 0.05, and that our STWCS-based framework reduces total annotation time by about 29% compared to typical adopted settings such as Simple Random Sampling (SRS) with a fixed sample size.

The remainder of the paper is organized as follows: Section 2 presents the background on accuracy estimation for KGs and introduces the GutBrainIE collection; Section 3 formalizes the employed NEL accuracy estimation framework, describing the definition of strata and clusters, the resulting estimator, and the considered cost model; Section 4 reports the results obtained on GutBrainIE, including corpus-level and strata-level accuracy estimates along with an efficiency analysis comparing our framework against a SRS baseline; finally, Section 5 draw conclusions and discusses directions for future work.

2. Background and Task Setting

2.1. Sampling and Accuracy Estimation in KGs

In sampling-based frameworks for KG accuracy estimation, the goal is to approximate the accuracy of the KG (proportion of correct triples over the total) by manually reviewing only a small subset of triples drawn with a certain sampling design [13, 15, 8]. The sampling design determines the statistical properties of the resulting estimator (e.g., variance, bias), the amount of human effort required, and, therefore, the overall annotation cost.

In this work, with *sampling* we mean drawing a subset of triples T_S from the full collection T and annotating only this subset to estimate the global accuracy $\mu(T)$. S is defined as the employed *sampling strategy*. We consider S as based on some sort of random sampling. Thus, the resulting estimator $\hat{\mu}(T_S)$ also exhibits some degree of randomness.

Following prior work on KG accuracy estimation [13, 15, 8], we associate an estimator $\hat{\mu}$ with a $(1-\alpha)$ CI, that is, an interval which contains the true accuracy $\mu(T)$ with probability approximately $(1-\alpha)$. Namely, the true accuracy $\mu(T)$ is bounded in:

$$[\hat{\mu} - \frac{\text{CI}}{2} \ , \ \hat{\mu} + \frac{\text{CI}}{2}]$$

By defining the *Margin of Error (MoE)* as the half-width of the CI, we can rewrite the bounds for $\mu(T)$ as

$$[\hat{\mu} - \text{MoE} \ , \ \hat{\mu} + \text{MoE}]$$

Since the MoE shrinks as the sample size $|T_S|$ grows, the goal is to design S to reach the target MoE $\leq \varepsilon$ while minimizing the number of manually annotated triples.

A straightforward choice for S is SRS, in which triples are selected uniformly at random from T without replacement. Under SRS, the sample mean provides an unbiased estimator of $\mu(T)$, and the corresponding CI can be derived using standard normal approximations [13]. However, SRS inherently scatters sampled triples across many unrelated documents, requiring annotators to repeatedly reconstruct the surrounding context for each triple. Because KGs are often large and highly heterogeneous –

even within specialized domains – SRS may, for example, present an annotator with a triple about an actor immediately followed by one about a geographic location. This contextual fragmentation creates inefficiency in both annotation time and cognitive burden. Indeed, as shown in works on KG accuracy estimation, sampling independent triples in most cases leads to increased annotation costs [13, 15].

To address this, Gao et al. [13] propose *Cluster Sampling (CS)*. The idea is to group triples that share the same contextual feature (e.g., the same subject entity) into *clusters*, and then sample clusters rather than individual triples for annotation. The rationale is that, once an annotator revises a certain triple, annotating additional triples drawn from the same cluster is more efficient than annotating triples from different clusters. Within this framework, two main cluster sampling designs are defined: Random Cluster Sampling (RCS), which samples clusters with a uniform random probability, and Weighted Cluster Sampling (WCS), which instead samples clusters with Probability Proportional to Size (PPS), i.e., proportional to the number of triples contained in each cluster. Both designs yield unbiased estimators of global accuracy [13].

CS designs still present a limitation: annotating every triple in selected clusters may be expensive, especially if considered clusters are large. To address this, Gao et al. [13] introduce Two-stage Weighted Cluster Sampling (TWCS). In the first stage, clusters are sampled with PPS (as in WCS); in the second stage, up to m triples are drawn without replacement from each sampled cluster and annotated. The resulting estimator $\hat{\mu}$ averages the within-cluster accuracies of sampled clusters and is proven to be unbiased for the overall accuracy $\mu(T)$. Intuitively, TWCS reduces annotation cost by placing an upper bound m on the number of triples annotated per cluster, while still leveraging the fact that sampling multiple triples within the same cluster is cheaper than sampling independent triples [13].

TWCS efficiency can be further improved by introducing *stratification*, i.e., partitioning the set of triples T into non-overlapping *strata* that are internally more homogeneous (e.g., in terms of expected accuracy, utility, or domain knowledge) than the entire set T as a whole [15, 8]. Stratification can be applied before or after clustering without any particular loss of generality. By merging stratification with TWCS, we obtain Stratified Two-stage Weighted Cluster Sampling (STWCS), a three-stage sampling design in which the first step is to sample a strata with Probability Proportional to Size With Replacement (PPS-WR), considering as size the number of triples contained in the strata. The last two steps apply TWCS as described above to the selected stratum. The rationale underlying STWCS is that, by building strata, triples with similar accuracy are grouped together; thus, variability within each stratum is lower than in the population as a whole. As a result, the estimator needs fewer annotated triples to reach the same MoE, and sampling can be focused on strata that have the most impact in reducing CI width, instead of being uniformly spread over the entire population, as in standard (non-stratified) TWCS [8].

2.2. Named Entity Linking for GutBrainIE

GutBrainIE is a large-scale biomedical corpus of PubMed abstracts focused on the gut-brain axis, which includes manual and automatic annotations for entities, concept-level links, and relations [14, 9, 10]. In recent years, the number of publications registered on PubMed on the gut-brain axis has been continuously increasing, with the yearly number of articles surpassing 2,000 in 2025.[1], making it increasingly difficult for field experts to stay up to date and extract relevant findings from this rapidly expanding body of literature. At the same time, gut-brain articles are characterized by complex, highly domain-specific terminology and semantics, which limit the effectiveness of generic biomedical IE models and hinder transfer learning from other biomedical corpora [16]. GutBrainIE is designed to foster the development and evaluation of gut–brain-specific IE systems that can support experts in automatically processing this literature [9],

GutBrainIE covers 13 entity types, including widely used biomedical categories (e.g., *anatomical location*, *bacteria*, *drug*) and entities specific to the gut-brain axis (e.g., *microbiome*, *dietary supplement*). To account for the frequent occurrence of experimental scenarios in documents, the corpus also includes specific categories related to medical experiments (e.g., *biomedical technique* and *statistical technique*).

[1]The value was obtained with the query `(gut brain axis[Title/Abstract]) OR (gut brain axis[Title/Abstract])` on PubMed.

On the relation side, GutBrainIE features 17 relation predicates, many of which are overloaded and can connect different combinations of entity types depending on context. This many-to-many design originates 55 distinct relation triples. For NEL, entity mentions are linked to concepts drawn from 6 standardized biomedical vocabularies, plus a custom ontology for unmatched mentions. By "*linking to a concept*" we mean assigning to the entity mention a Uniform Resource Identifier (URI) that resolves to the corresponding concept in a reference resource.

To ensure consistency across the corpus, GutBrainIE defines, for each entity type, a priority order over reference resources so that when the same concept is available in multiple resources, the selected concept is always drawn from the highest-priority one. The full list of entity types and their associated resources is reported in Table 1.

Table 1
Biomedical resources used for concept-level linking of entity mentions. For each entity label, the table lists the reference resources ordered accordingly to the priority considered in NEL annotations. *GBIE* indicates our custom-defined vocabulary.

Entity Label	Linked Vocabularies
Anatomical Location	UMLS, NCIT, GBIE
Animal	UMLS, NCIT, NCBITaxon, GBIE
Bacteria	UMLS, NCIT, NCBITaxon, MESH, OMIT, GBIE
Biomedical Technique	UMLS, NCIT, OMIT, NCBITaxon, GBIE
Chemical	UMLS, NCIT, CHEBI, OMIT, GBIE
Dietary Supplement	NCIT, UMLS, CHEBI, NCBITaxon, OMIT, MESH, GBIE
DDF	UMLS, NCIT, OMIT, NCBITaxon, GBIE
Drug	UMLS, NCIT, CHEBI, OMIT, NCBITaxon, GBIE
Food	UMLS, NCIT, GBIE
Gene	UMLS, NCIT, OMIT, CHEBI, GBIE
Human	UMLS, MESH, GBIE
Microbiome	UMLS, NCIT, NCBITaxon, GBIE
Statistical Technique	UMLS, NCIT, GBIE

The GutBrainIE collection includes a total of 1'647 documents and is organized into four folds that reflect annotation quality and annotator expertise. Platinum and Gold folds contain annotations curated by domain experts (399 documents), Silver is annotated by trained laypersons (499 documents), while Bronze is automatically annotated without subsequent manual revision (749 documents). Considering manually annotated documents only, these present a high density of annotations, featuring, on average, 29.46 entities and 15.47 relations per document.

NEL annotations are automatically generated by a linking system that combines heuristic normalization rules with lexical and similarity matching against the reference resources. For that reason, NEL annotations are only for the expert-curated folds (Platinum and Gold), since layperson and automatic annotations lack the precision required for reliable automatic concept mapping. In total, this pipeline produces more than 11,000 concept-level links. Manually revising all of them would be extremely costly and time-consuming. Therefore, we propose a statistically reliable framework to estimate overall NEL accuracy while limiting manual revision costs and supporting targeted manual corrections instead of comprehensive re-annotation.

3. Accuracy Estimation for NEL Annotations

We follow previous work on sampling-based accuracy estimation for KGs (c.f., Section 2.1), where the goal is to estimate the overall accuracy of KG triples while minimizing the number of manual annotations and providing statistical guarantees about the resulting estimates. Without loss of generality, we can view each NEL annotation in our setting as a triple in a KG, where the mention (text span + entity label) acts as subject, the concept URI as object, and a generic *hasConcept* predicate connects them. Therefore,

we can directly apply the same methodologies used for KG accuracy estimation to NEL annotations.

Let T be the set of all such triples. We define a correctness indicator function:

$$\mathbb{1}_T(t) \in \{0, 1\},$$

where $\mathbb{1}_T(t) = 1$ if the triple is judged correct by an expert annotator, 0 otherwise. A triple is considered incorrect if the entity mention is linked to a wrong concept or if the assigned concept is overly generic, i.e., a more specific and appropriate one exists in the reference resources. Our goal is to estimate the mean correctness of all such triples:

$$\mu(T) = \frac{1}{|T|} \sum_{t \in T} \mathbb{1}_T(t), \tag{1}$$

using manual annotations collected only for a subset T_S of T. Specifically, a sampling-based evaluation framework replaces $\mu(T)$ with an estimator $\hat{\mu}$ computed over a sample $T_S \subset T$ drawn according to some sampling strategy S. To be meaningful, the estimator should be (approximately) unbiased, i.e., $E[\hat{\mu}] = \mu(T)$, and accompanied by a $(1 - \alpha)$ CI with controlled MoE, defined as half the CI width. Unbiasedness is crucial because it guarantees that, on average, the accuracy estimated from the sampled triples T_S reflects the true accuracy $\mu(T)$ of the entire set T. Moreover, this makes the associated CIs reliable in quantifying how well the estimated accuracy on T_S reflects the corpus-level accuracy μ_T.

3.1. Sampling Design

In this work, we adopt STWCS as our sampling design and adapt it to NEL triples of the GutBrainIE collection. To do that, we need to define: (i) *strata*, which partition triples into non-overlapping groups, typically to reduce variance and incorporate domain knowledge, and (ii) *clusters*, which group together triples that share a contextual feature.

Strata. We assign triples to one of 5 strata defined in a fully deterministic way based on their entity labels:

1. **DDF** (Disease, Disorder, or Finding): 4,015 triples;
2. **Microbiome + Bacteria**: 2,030 triples;
3. **Human + Animal + Anatomical Location**: 2,078 triples;
4. **Chemical + Gene**: 1,487 triples;
5. **Drug + Dietary Supplement + Food + Biomedical Technique + Statistical Technique**: 1,484 triples.

The definition of these strata took into account semantic coherence and uniform cardinality. By semantic coherence we mean grouping entity labels that tend to appear in similar contexts. For instance, *Microbiome* and *Bacteria* are grouped since bacteria compose the microbiome; *Human*, *Animal*, and *Anatomical Location* are placed in the same stratum since they all describe biological subjects and their body parts; *Chemical* and *Gene* are aggregated since chemicals can be considered genes and viceversa depending on the context. We validated our stratification design by asking the biomedical experts involved in the GutBrainIE annotation efforts (thus, highly familiar with the collection and its annotation schema) to review the proposed groupings and they confirmed that these were consistent with the domain semantics.

Other stratification strategies would have been possible, such as adopting single-label strata (i.e., one stratum per entity category) or grouping entities by ontology resource provenance. The first strategy would be severely sub-optimal, since increasing the number of strata tends to generate very small partitions, which can negatively impact the minimization of the manual annotation efforts – i.e., the number of manual annotations required to reach the target MoE. The latter strategy, instead, has to be avoided since, in this context, there is no correlation between resource provenance and

expected accuracy. Therefore, this approach would likely yield high within-stratum variance, making the sampling process inefficient.

Looking at the cardinality of defined strata, it's immediate to notice that DDF is considerably higher in size compared to others. This is expected given the nature of the GutBrainIE collection, whose primary goal is to foster the development of IE systems that can automatically extract knowledge from scientific literature about the gut-brain interplay. In this context, a core interest is understanding the impact of various factors (microbiome, chemicals, drugs, foods, etc.) on diseases, disorders, and findings. This naturally makes DDF the most prevalent category. Moreover, given that DDF is treated as a single entity category in GutBrainIE, further dividing it into sub-categories was not possible. Introducing an automatically generated sub-division would have been highly unreliable, as also medical specialists often find it difficult to consistently separate diseases, disorders, and findings. On the other hand, manually creating this sub-division would have been extremely time-consuming and laborious, contradicting the purpose of this accuracy estimation framework in minimizing manual curation efforts. The resulting stratification still achieves a reasonable balance across the remaining strata while putting a bit more focus on the DDF stratum, reflecting the intended focus of the collection.

Clusters. Within each stratum, we define cluster groups as all triples whose mention text spans normalize to the same surface form. Let ts_f be the mention text span of a triple t, and let σ_j indicate the normalized surface form associated with cluster j. Then, we can define a generic cluster j as:

$$\text{Cluster}_j = \{t \in T \mid \text{normalize}(ts_f) = \sigma_j\} \tag{2}$$

where normalization lowercases the text span and removes extra whitespaces. This design makes annotators reviewing multiple occurrences of the same surface form in sequence, thus reducing context switches and improving annotation efficiency. Moreover, combined with the way of annotating incorrect links described in Section 3, this clustering approach enables downstream corrections at the cluster level, for example by correcting all at once mentions in the same cluster that were mapped to an over-general or wrongly assigned concept.

Table 2 shows the number of triples and unique clusters for each stratum. The higher weight of the *DDF* stratum reflects the fact that this label actually groups three categories (diseases, disorders, and findings), which occur much more frequently than other entity types in the PubMed documents composing the GutBrainIE collection.

STWCS procedure. Sampling is done iteratively and consists of three steps:

1. **Stratum selection**: sample a stratum h with PPS-WR, where size is the total number of triples $M[h]$ in the stratum.
2. **Cluster selection**: within the selected stratum, sample a cluster j with Probability Proportional to Size Without Replacement (PPS-WOR), considering as size the number of triple in the cluster ($|\text{Cluster}_j|$).
3. **Triples annotation**: within the selected cluster, annotate up to m triples (or all triples if the cluster is smaller).

In step (3) we fix $m = 5$, proven by [13] to be a near-optimal choice that minimizes evaluation cost for TWCS across different KGs while preserving its efficiency gains over SRS. After each iteration, we update the accuracy estimates and associated CI, and we continue sampling and annotating until the target MoE is reached. It's worth noting that both stratification and clustering are defined independently of NEL outputs. This is crucial since we are evaluating NEL itself and, therefore, cannot rely on its (a priori unknown) quality in our sampling strategy. This is reinforced by the fact that, as discussed in Section 2.2, NEL annotations are produced automatically. Our design choices avoid circularity between what is being evaluated and how the evaluation sample is constructed.

3.2. Estimator and Confidence Interval

We first recall the estimator used by TWCS within a single (unstratified) set of triples T. Let n be the number of sampled clusters and, for the k-th sampled cluster, let $\mu_{I,k}$ indicate the mean accuracy of the (at most m) triples annotated in that cluster. The TWCS estimator of the overall accuracy in this setting is given by:

$$\hat{\mu}_{w,m} = \frac{1}{n} \sum_{k=1}^{n} \mu_{I,k} \tag{3}$$

which is proven to be an unbiased estimator of the true accuracy under the TWCS sampling design [13].

STWCS extends TWCS by applying it independently within each stratum and then aggregating the resulting estimates. Let $M[h]$ be the total number of triples in stratum h, $M = \sum_h M[h]$ the total number of triples, and $W_h = M[h]/M$ the weight of stratum h. Let $\hat{\mu}_{w,m,h}$ be the TWCS estimator computed within stratum h. The STWCS estimator of the global accuracy is then:

$$\hat{\mu}_{ss} = \sum_h W_h \cdot \hat{\mu}_{w,m,h} \tag{4}$$

and the $(1-\alpha)$ CI for $\hat{\mu}_{ss}$ can be written as:

$$\hat{\mu}_{ss} \pm z_{\alpha/2} \sqrt{\sum_h W_h^2 \cdot \mathrm{Var}(\hat{\mu}_{w,m,h})},$$

where the variance term is estimated from the sampled clusters in each stratum. As shown in [13], the STWCS estimator in Eq. 4 is unbiased for the true corpus-level accuracy $\mu(T)$ under the sampling design described in Section 3.1.

3.3. Cost Model and Context-Switching

To assess efficiency, we introduce a simple cost model that captures the impact of context switches on annotation time. We define a *context switch* as any transition between two consecutively annotated triples belonging to different surface-form clusters. Intuitively, switching clusters requires the annotator to adapt to a new textual and conceptual context, increasing annotation time.

Let n_t be the total number of annotated triples in a sample T_S, and let n_{sw} be the number of context switches occurred while annotating T_S. We model the cost of annotating T_S as

$$\mathrm{cost}(T_S) = c_t \cdot (n_t - n_{\mathrm{sw}}) + c_s \cdot n_{\mathrm{sw}} \tag{5}$$

where c_t is the per-triple cost without context switch and c_s is the per-triple cost when a context switch occurs.

In our analysis, we measure cost in terms of wall-clock time. Let t_{base} denote the mean annotation time without context switch and $\Delta t_{\mathrm{switch}}$ the additional time introduced by switching context. The associated time-cost model can be written as:

$$\mathrm{time}(T_S) = n_t \cdot t_{\mathrm{base}} + n_{\mathrm{sw}} \cdot \Delta t_{\mathrm{switch}} \tag{6}$$

In the following, we will discuss wall-clock time results measured on the actual annotation campaign.

3.4. Annotation Interface

To collect correctness annotations on sampled triples, we implemented a custom web application using *streamlit*.[2] This application supports the annotation workflow described in Section 3, allowing annotators to decide, for each triple, whether the link between mention and concept is correct, incorrect

[2] https://streamlit.io/

because mapped to an overly general concept, or incorrect because the linked concept is wrong. The application also records the wall-clock time spent in annotating each triple, which is later used to compute values for the cost model illustrated in Section 3.1. Moreover, our application supports collaborative work by allowing annotators to export their annotations as a JSON file that can be used to resume their work or extend the one done by other experts.

Input Format. The input expected by our web application is a JSONL file containing the triples to be annotated and, optionally, a JSON file with annotations from a previous sessions (either by the same or another annotator). To simplify the setup and avoid having to implement real-time iterative sampling, we generated a static sample batch before starting the actual annotation process. To do that, we simulated an annotation round by sampling triples with our STWCS design (see Section 3.1), fixing the random seed to 42 for reproducibility. To ensure that the resulting batch was large enough to reach the target MoE, we assigned synthetic labels for correctness/incorrectness under a worst-case assumption for convergence, i.e., simulating an accuracy around 0.5 by alternating triples as correct and incorrect. This procedure created a static batch comprising 7,517 triples. In the produced JSONL file, each line corresponds to a single triple, and lines are ordered according to the STWCS sampling order. The annotation interface presents triples in the order they are reported in this file, thus preserving the strata and clusters structure imposed by STWCS, which aims to minimize context switches.

MoE Convergence Monitoring. To notify annotators when the number of reviewed triples was sufficient to reach MoE $\leq \varepsilon = 0.05$, we implemented a separate monitoring script in Python. This script is launched before starting an annotation session and then keeps running on the background. Every time the annotator finishes evaluating all triples belonging to a sampled cluster, the script automatically recomputes the estimated accuracy and associated MoE. If the computed MoE goes below the target threshold ε, the script triggers a desktop notification (using the *plyer* Python package[3]), telling the annotator that convergence has been reached and further annotations are not needed for the global accuracy estimate.

User Interface. The annotation interface is shown in Figure 1. Panel (1) allows annotators to load a saved JSON file containing previous annotations, enabling them to resume work from another annotator or from an earlier annotation session. Panel (2) summarizes progress, displaying the index of the current triple, the number of annotated triples over the total, and providing navigation controls. By double-clicking on the "*Next*" arrow, annotators can open a dialogue box and directly jump to a specific triple index, which is useful when continuing previous sessions. This panel also includes an "*Export*" button that allows annotators to save all annotations produced so far to a JSON file with a custom name and to a custom location. Notice that the application still automatically saves annotations to a JSON file in a local cache folder every time the user moves to a different triple. Panel (3) shows the full abstract text with the current mention highlighted, giving annotators all necessary local context. Panels (4) and (5) display details about the entity and the linked concept. Specifically, (4) includes the mention text, label, location (title/abstract) and offsets in the location (in terms of number of characters, including whitespaces), while (5) reports the URI along with associated names and definitions retrieved from the reference resource. Finally, panel (6) contains the annotation controls and a button to submit the rating and proceed to the next triple.

4. Results

4.1. Annotation Campaign and Sampling Outcome

The expert folds of GutBrainIE contain a total of 11,184 NEL triples organized into 4,116 global surface-form clusters across the five strata described in Section 3.1. Using the STWCS design, we iteratively

[3]https://github.com/kivy/plyer

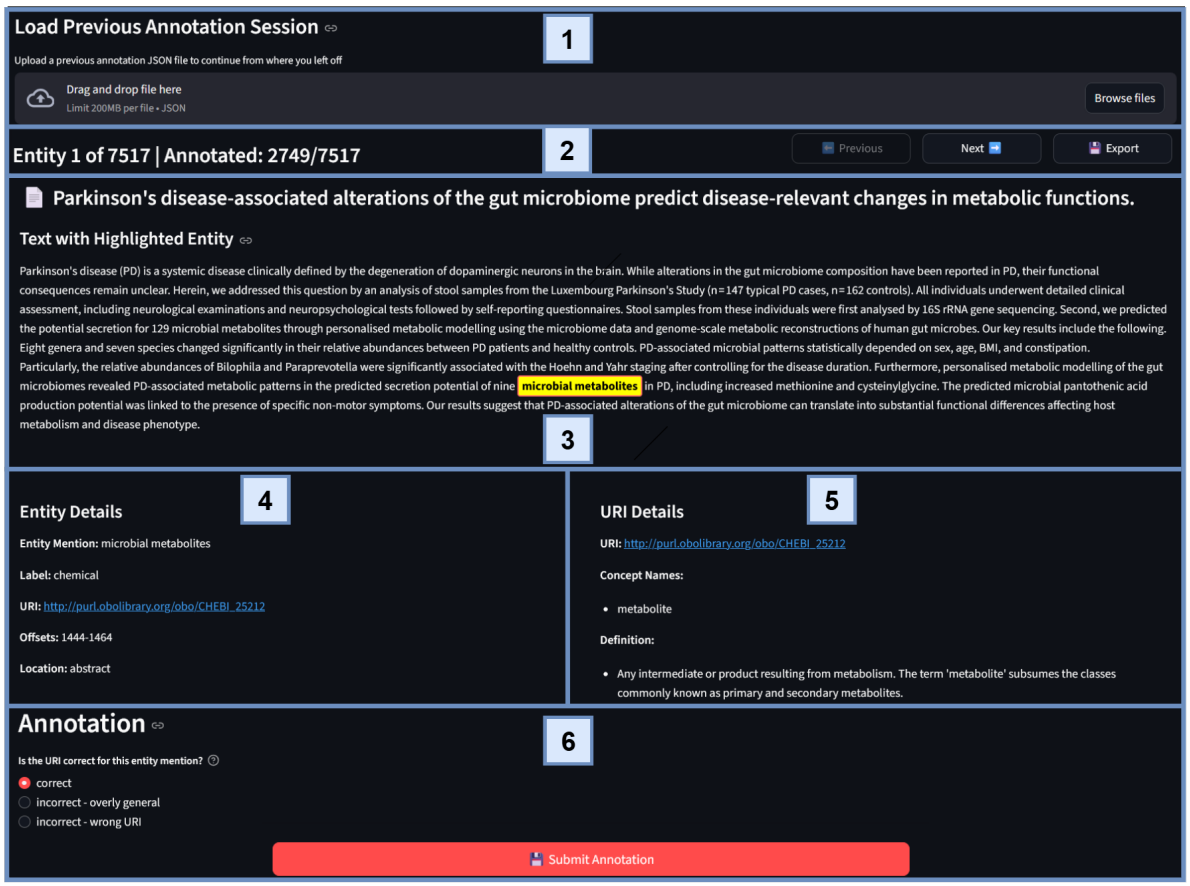

Figure 1: Annotation interface used for the revision of NEL triples.

sampled and annotated triples until reaching a target MoE (ε) of 0.05 for the overall accuracy estimate.

To carry out the annotation process, we recruited an expert annotator with a strong biomedical background and prior experience in annotation campaigns. Each sampled NEL annotation was therefore reviewed by a single assessor. While multiple independent annotations per instance would increase reliability, our revision task was limited to verifying the assigned concept and labelling it as *correct*, *wrong*, or *overly general*. The annotator was provided with the complete list of concepts included in the GutBrainIE collection, making this verification task straightforward for experienced assessors.

Convergence was obtained after annotating 2,749 triples (24.6% of all triples), belonging to 1,044 unique clusters. Across subsequent annotations, we observed 1,050 context switches and 1,698 no-switch transitions (i.e., transitions within the same cluster).

The resulting overall NEL accuracy estimate is

$$\hat{\mu}_{ss} = 0.915 \pm 0.0473,$$

corresponding to a $(1-\alpha)$ CI of $[0.868, 0.963]$. The resulting MoE is 0.047, thus below the target threshold $\varepsilon = 0.05$.

4.2. Accuracy Estimates by Stratum

STWCS also produces per-stratum accuracy estimates. The number of annotated clusters and triples for each stratum is reported in Table 2, while the estimated accuracy $\hat{\mu}$, the corresponding MoE, and the resulting $(1-\alpha)$ CI for each stratum are reported in Table 3.

While strata-level intervals are wider than the overall CI (as expected given the reduced number of triples annotated per group), they provide useful insights into how NEL accuracy varies across different

Table 2
Distribution of weights, clusters, and triples across strata. For each stratum, the *Clusters* and *Triples* columns report the number of sampled (i.e., annotated) items over the total available.

Stratum	Weight	Clusters (Annotated / Total)	Triples (Annotated / Total)
DDF	0.3670	366 / 1,321	1,013 / 4,105
Microbiome + Bacteria	0.1815	197 / 514	513 / 2,030
Human + Animal + Anatomical Location	0.1858	209 / 665	559 / 2,078
Chemical + Gene	0.1330	147 / 802	381 / 1,487
Drug + Dietary Supplement + Food + Biomedical/Statistical Technique	0.1327	128 / 814	283 / 1,484

Table 3
Strata-level NEL accuracy estimates, reporting for each stratum the estimated $\hat{\mu}$ and the corresponding margin of error (MoE).

Stratum	$\hat{\mu}$	MoE
DDF	0.883	0.093
Microbiome + Bacteria	0.979	0.108
Human + Animal + Anatomical Location	0.976	0.060
Chemical + Gene	0.851	0.087
Drug + Dietary Supplement + Food + Biomedical/Statistical Technique	0.898	0.154

groups of entity types. Achieving narrower strata-level CIs would require additional annotations, which our framework supports by continuing sampling in specific strata.

Overall, if we focus on the lower bounds of the strata-level CIs ($\hat{\mu} - \text{MoE}$) as conservative estimates of accuracy, we observe for all strata an accuracy estimate of at least 0.74, indicating satisfactory NEL quality across all considered entity groups. In particular, *Microbiome + Bacteria* and *Human + Animal + Anatomical Location* strata obtain lower bounds above 0.85, highlighting high NEL reliability for these entity labels. These strata-level estimates can be leveraged as guidance for prioritizing downstream targeted curation on strata showing lower accuracy.

4.3. Efficiency Analysis vs Simple Random Sampling

To quantify the efficiency gains coming from our STWCS sampling design, we analyze annotation times and compare them to a SRS baseline. About the latter, since re-running the entire annotation campaign under SRS would be unnecessarily costly, we estimate it via simulation. Specifically, we keep fixed the same 2,749 triples annotated under STWCS and simulate what would happen if they were drawn and annotated with SRS instead (i.e., with uniform random probability). Notice that an actual SRS-based sampling design would generally select a different set of triples and, potentially, require annotating a different number of triples, with a different number of context switches, to reach the target MoE.

Operationally, each SRS simulation produces a random permutation of the triples, which we use to count the number of transitions with and without context switches. We repeat this process 1,000 times and then bootstrap the resulting transition counts to reduce the impact of rare (i.e., outlier) SRS permutations.

The total time spent by experts to annotate the 2,749 sampled triples under STWCS was 797 minutes (13 hours and 17 minutes). From the logs collected during the annotation campaign, we estimate:

- mean annotation time without context switch (t_{base}) of 0.22 minutes (12.97 seconds);
- mean annotation time with context switch ($t_{\text{base}} + \Delta t_{\text{switch}}$) of 0.41 minutes (24.57 seconds),

From that, we can derive that the additional cost per context switch (Δt_{switch}) is 0.19 minutes (11.59 seconds). Therefore, switching context introduces a slowdown factor of $0.41/0.22 \approx 1.89$.

To provide a realistic estimate of the annotation cost that SRS would have, we estimate the total annotation time under SRS using the same t_{base} and Δt_{switch} measured above. We bootstrapped the distribution resulting from 1,000 SRS simulations and obtained a mean number of context switches of

$2,745$, a mean number of no-switch transitions of 2.84, and a mean total annotation time of 1,124.57 minutes (18 hours and 44 minutes).

Compared to this SRS baseline, STWCS saves approximately 327.6 minutes (5 hours and 28 minutes) of expert time at fixed sample size. The corresponding efficiency ratio is $\frac{\text{time}_{\text{STWCS}}}{\text{time}_{\text{SRS}}} \approx 0.71$, indicating that our sampling framework achieves the same statistical precision with about 29% less annotation time. This demonstrates that applying STWCS to NEL accuracy estimation gives efficiency gains over SRS that are aligned to those reported by [13, 8] for KG accuracy estimation, even though our setting focuses on NEL triples rather than KG triples.

4.4. Generalization and Applicability

Although our evaluation is conducted solely on the GutBrainIE collection, the proposed framework is not tied to a specific dataset or setting. Indeed, the STWCS estimator and the MoE-based stopping criterion only require: (i) a population of annotations to be evaluated and (ii) a strategy for partitioning this population into strata and clusters. In this work, strata are based on entity categories, while clusters rely on normalized surface forms. However, these design choices can be easily adapted to different benchmarks and contexts with minimal effort.

In general, strata should be defined to group items expected to show similar error patterns, since reducing within-stratum variance directly improves sampling efficiency. Features that can be leveraged in the stratification design include: entity types/categories when available (e.g., PER/ORG/LOC in general-purpose NEL), the ontology/resource provenance of the linked concept (e.g., UMLS vs other resources), and mention-level features correlated with difficulty or ambiguity (e.g., mention text span length or frequency). Remarkably, the choice of the stratification strategy does not affect the validity of the STWCS estimator. Namely, even if the assumed correlation is weak or totally absent, the estimator remains unbiased. What changes is the variance of the estimate and, consequently, the number of manual verifications required to reach the target MoE.

For what concerns clusters, they should be defined to minimize manual annotation efforts. In general, this is obtained by exploiting redundancy in the considered population. A simple and effective approach is to group mentions by lexical similarity of their text spans, since similar mentions tend to share the same context and error patterns. Lexical similarity can be implemented through different normalization functions (e.g., casing, punctuation, non-latin letter variants) or by using standard Natural Language Processing (NLP) toolkits (e.g., NLTK or spaCy). When surface forms are highly variable, clustering can be extended to consider additional cues, such as the local mention context (i.e., surrounding tokens), or document-level metadata.

5. Conclusion and Future Directions

We presented a sampling-based framework for estimating NEL accuracy in the GutBrainIE collection under statistical guarantees and limited annotation budget. By framing NEL accuracy estimation as a constrained optimization problem and adapting the STWCS design to NEL triples, we were able to estimate a corpus-level accuracy of 0.915 ± 0.0473 by annotating less than 25% of all triples and reducing the total annotation time by about 29% compared to a simulated SRS baseline, demonstrating the effectiveness of our sampling design focused on minimizing context switches. Moreover, we were able to obtain strata-level accuracy estimates representing NEL accuracy across groups of entity types.

Our design uses only NEL-independent features (entity labels and surface forms) to define strata and clusters, avoiding circularity between the evaluation target and the sampling approach. Moreover, it gives the possibility to continue sampling in specific strata to tighten local CIs (i.e., strata-level CIs) when needed. Finally, given that the proposed framework is not specific to GutBrainIE, it can be applied to other NEL benchmarks and IE pipelines that require scalable and statistically robust accuracy assessment under limited expert budgets.

Future work might employ multiple annotators per triple and aggregate their labels (e.g., via majority voting), providing more reliable annotations on the correctness of linked concepts. Moreover, stratifica-

tion could incorporate downstream impact, for instance, by prioritizing concepts that participate in many relation instances or have high centrality in the induced KG.

Acknowledgments

This project has received funding from the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.2 and Grammarly in order to: Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] W. Shen, J. Wang, J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Transactions on Knowledge and Data Engineering 27 (2014) 443–460.

[2] T. Al-Moslmi, M. G. Ocaña, A. L. Opdahl, C. Veres, Named entity extraction for knowledge graphs: A literature overview, IEEE access 8 (2020) 32862–32881.

[3] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of biomedical informatics 137 (2023) 104252.

[4] S. R. Jonnalagadda, P. Goyal, M. D. Huffman, Automating data extraction in systematic reviews: a systematic review, Systematic reviews 4 (2015) 78.

[5] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, Journal of biomedical informatics 77 (2018) 34–49.

[6] J. Stewart, J. Lu, A. Goudie, G. Arendts, S. A. Meka, S. Freeman, K. Walker, P. Sprivulis, F. Sanfilippo, M. Bennamoun, et al., Applications of natural language processing at emergency department triage: A narrative review, Plos one 18 (2023) e0279953.

[7] Y. Qi, W. Zheng, L. Hong, L. Zou, Evaluating knowledge graph accuracy powered by optimized human-machine collaboration, in: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, 2022, pp. 1368–1378.

[8] S. Marchesin, G. Silvello, O. Alonso, Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 12, 2024, pp. 105–114.

[9] M. Martinelli, G. Silvello, V. Bonato, G. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, et al., Overview of GutBrainIE@ CLEF 2025: Gut-Brain Interplay Information Extraction, in: CEUR Workshop Proceedings, volume 4038, CEUR-WS, 2025, pp. 65–98.

[10] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, et al., Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 173–198.

[11] M. Carabotti, A. Scirocco, M. A. Maselli, C. Severi, The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems, Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology 28 (2015) 203.

[12] J. Appleton, The gut-brain axis: influence of microbiota on mood and mental health, Integrative Medicine: A Clinician's Journal 17 (2018) 28.

[13] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, J. Yang, Efficient knowledge graph accuracy evaluation, arXiv preprint arXiv:1907.09657 (2019).

[14] M. Martinelli, S. Marchesin, V. Bonato, G. M. D. Nunzio, N. Ferro, O. Irrera, L. Menotti, F. Vezzani, G. Silvello, A Domain-Specific Curated Benchmark for Entity and Document-Level Relation Extraction, 2026. URL: https://arxiv.org/abs/2602.04320. arXiv:2602.04320.

[15] S. Marchesin, G. Silvello, Efficient and reliable estimation of knowledge graph accuracy, Proceedings of the VLDB Endowment 17 (2024) 2392–2403.

[16] M. Jafari, X. Tao, P. Barua, R.-S. Tan, U. R. Acharya, Application of transfer learning for biomedical signals: a comprehensive review of the last decade (2014–2024), Information Fusion 118 (2025) 102982.