LongBEL: Long-Context and Document-Consistent Biomedical Entity Linking

Anonymous ACL submission

Abstract

Biomedical entity linking maps textual mentions to concepts in structured knowledge bases such as UMLS or SNOMED CT. Most existing systems link each mention independently, using only the mention or its surrounding sentence. This ignores dependencies between mentions in the same document and can lead to inconsistent predictions, especially when the same concept appears under different surface forms. We introduce **LongBEL**, a document-level generative framework that combines full-document context with a memory of previous predictions. To make this memory robust, LongBEL is trained with cross-validated predictions rather than gold labels, reducing the mismatch between training and inference and limiting cascading errors. Experiments on five biomedical benchmarks across English, French, and Spanish show that LongBEL improves over sentence-level generative baselines, with the largest gains on datasets where concepts frequently recur within documents. An ensemble of local, global, and memory-based variants achieves the best results across all benchmarks. Further analysis shows that the largest gains occur on recurring concepts, suggesting that LongBEL mainly improves document-level consistency rather than isolated mention disambiguation.

1 Introduction

Biomedical Entity Linking (BEL) maps textual mentions in biomedical documents to concepts in structured knowledge bases (KBs) such as the Unified Medical Language System (UMLS) or SNOMED CT (Bodenreider, 2004; Donnelly, 2006). This normalization step is essential for biomedical information extraction, as it resolves lexical variation and allows mentions referring to the same concept to be aggregated (Krauthammer and Nenadic, 2004).

Recent neural models have substantially improved BEL performance, especially through contextualized representations. However, most state-of-the-art BEL systems still link mentions independently, using only local context (Chen et al., 2026). This is limiting in biomedical documents, where concepts can reappear through abbreviations, synonyms, or partial expressions, and where earlier mentions can provide useful semantic cues. Ignoring these document-level dependencies can lead to inconsistent predictions within the same document (Schwartz and Hearst, 2002).

Document-level consistency remains underexplored in current BEL systems. Generative BEL models cast linking as a sequence-to-sequence task and can handle large candidate spaces (Kim et al., 2025; Yuan et al., 2022a), but they are usually applied with local context only. As illustrated in Figure 1, useful evidence may be distributed across the document, either through repeated concepts or related clinical cues.

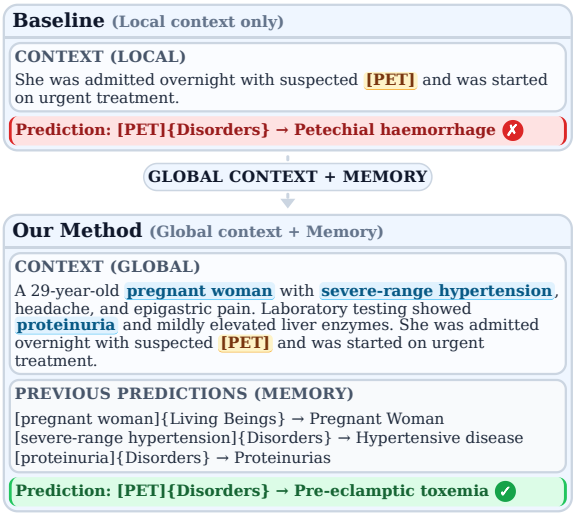

Figure 1: Illustrative example of document-level context and memory. **Top:** Restricted to local context, a standard model mislinks *[PET]* to *Petechial haemorrhage*. **Bottom:** By using global clinical context and previous predictions, LongBEL correctly links *[PET]* to *Pre-eclamptic toxemia*.

We propose **LongBEL**, a document-level generative framework for BEL. LongBEL combines full-document context with a robust prediction memory that stores previous linking decisions and reuses them to improve consistency across a document.

Our main contributions are:

- We design a robust prediction-memory mechanism that reuses previous linking decisions while limiting cascading errors.
- We combine memory with long-context generation and ensemble decoding, obtaining the best results on five BEL benchmarks across English, French, and Spanish, while comparing against reproduced techniques from recent BEL work.
- We provide a detailed analysis showing that document-level information helps most for recurring concepts, and that this effect varies across datasets.

All code, preprocessing scripts, and evaluation scripts needed to reproduce the experiments are available in an anonymous GitHub repository. Trained model checkpoints, processed datasets, and the resources required to run inference are available in an anonymous Hugging Face repository.

2 Related Work

2.1 Context-Free BEL

Rule-Based and Lexical Systems. Early BEL systems such as MetaMap (Aronson and Lang, 2010), cTAKES (Savova et al., 2010), and scispaCy (Neumann et al., 2019) rely on lexical matching against biomedical dictionaries. These systems are efficient and easy to deploy, but they struggle to generalize to unseen variations and consistently fail when encountering ambiguous mentions.

Self-Supervised Bi-Encoders. Dense bi-encoder models such as SapBERT (Liu et al., 2021a), CODER (Yuan et al., 2022b), and BERGAMOT (Sakhovskiy et al., 2024) learn biomedical concept representations from ontology synonyms using contrastive learning. These models capture lexical variation better than string-matching systems, but they usually encode mentions independently and lack the contextual grounding necessary to resolve ambiguities, such as acronyms that refer to distinct concepts in different clinical settings.

2.2 Local-Context BEL

Recent BEL models incorporate the sentence or neighboring context around each mention.

Contextualized Bi-encoders. BLINK (Wu et al., 2020) introduced dense retrieval for entity linking by encoding mentions with local context and comparing them to entity descriptions. ArboEL (Agarwal et al., 2022) adapted this approach to the biomedical domain and added a structure-aware reranking step to improve consistency across related mentions.

Generative Models. Generative approaches build on GENRE (Cao et al., 2021), which formulates entity linking as constrained autoregressive generation. Biomedical variants such as GenBioEL (Yuan et al., 2022a) and ANGEL (Kim et al., 2025) adapt constrained decoding to biomedical ontologies. These models can search large candidate spaces efficiently, but they are usually applied at the sentence level and remain limited to local context.

Large Language Models for BEL. Large language models have also been used for BEL, most often as rerankers, prompting-based disambiguators, or data generators (Ding et al., 2024; Xie et al., 2024; Wang et al., 2025; Ding et al., 2025; Ai et al., 2025; Shlyk et al., 2024; Lin et al., 2025; Vollmers et al., 2025; Xin et al., 2025; Remaki et al., 2026; Borchert et al., 2024). However, these systems usually remain auxiliary to retrieval-based pipelines or sentence-level linkers, and do not directly enforce document-level consistency during decoding.

2.3 Document-Level Modeling and Consistency

Document-Level Consistency. Document-level consistency has been studied in general-domain entity linking through global coherence objectives and collective inference (Ratinov et al., 2011; Ganea and Hofmann, 2017). In BEL, ArboEL (Agarwal et al., 2022) models mention to mention structure to encourage consistency. However, these consistency-aware models remain fundamentally structured as retrieval pipelines: cross-mention agreement is encouraged indirectly via representation learning or collective inference, rather than strictly enforced through document-level autoregressive decoding.

Long-Context Autoregressive Decoding. Recent open-weight language models have achieved

dramatic increases in context capacity (Grattafiori et al., 2024), making document-level processing increasingly viable. Related work has used memory mechanisms to track entities across long narratives (Zhu et al., 2025), but such mechanisms have not been directly explored for constrained BEL.

2.4 Positioning of LongBEL

LongBEL combines long-context generative modeling with prediction memory for BEL. Unlike sentence-level generative systems, it conditions each prediction on the full document. Unlike retrieval-based consistency methods, it integrates previous linking decisions directly into the autoregressive decoding process. This allows LongBEL to improve document-level consistency without relying on external graph clustering or reranking.

3 Our Methods

3.1 Problem Statement

Let $\mathcal{E}$ be the set of candidate concepts in a KB such as UMLS. Each concept $e \in \mathcal{E}$ has a semantic group g_e (e.g., *Disorders*) and a set of synonyms S_e. Given a document c (the *global context*) containing an ordered sequence of mentions $\mathcal{M} = (m_1, m_2, \ldots, m_n)$, where each mention m_t is associated with a semantic group g_{m_t}, the goal is to map each mention m_t to its target concept $e_t \in \mathcal{E}$. For example, in Figure 1, the mention *PET* belongs to the *Disorders* semantic group and should be linked to the concept C0032914, whose synonyms include *Pre-eclamptic toxemia*.

3.2 Document-level Autoregressive Formulation with Memory

We frame this entity linking task as a sequence-to-sequence problem, extending the autoregressive formulation of Cao et al. (2021) to a document-level setting. To disambiguate the t-th mention m_t, our model relies not only on the global context c, but also on a *memory* $\mathcal{H}_t$ of concepts predicted earlier in the document.

Memory Mechanism. Concepts within the same document are often related or repeated. To use this information, LongBEL keeps a memory of previous linking decisions. At step t, the memory is defined as:

$$\mathcal{H}_t = \{(m_1, e_1), \ldots, (m_{t-1}, e_{t-1})\}.$$

For example, before linking the mention *PET* in Figure 1, the memory may already contain earlier predictions such as (*pregnant woman*, *Pregnant Woman*), (*severe-range hypertension*, *Hypertensive disease*), (*proteinuria*, *Proteinurias*).

Input Format. For the t-th mention, the input sequence x_t is constructed by concatenating three components: (i) the full document c, (ii) the memory $\mathcal{H}_t$ (formatted as a list of previous concepts), and (iii) the current target mention m_t alongside its semantic group g_{m_t}. In the same example, the target mention is *PET*, with semantic group *Disorders*, and the input also includes the full clinical paragraph and the previous memory entries.

Dataset Construction and Target Selection. For each mention m_t, the target text y_t is chosen from the synonym set S_{e_t} of its gold concept. Instead of always using the canonical concept name, we followed SynCABEL (Remaki et al., 2026) and select an unambiguous synonym that best matches the mention surface form. Concretely, we first remove synonyms that are associated with multiple concept identifiers within the same semantic group. The remaining synonyms are then ranked using TF-IDF similarity with the mention, and the highest-scoring synonym is used as the target string. This yields target strings that remain close to the observed mention while preserving a deterministic mapping to a unique concept identifier. The same unambiguous synonym inventory is used at inference time to build the constrained decoding trie, so each completed decoded string maps to exactly one KB concept.

Training Objective. The model is trained with teacher forcing to generate the target concept text y_t conditioned on the input x_t:

$$p_\theta(y_t \mid x_t) = \prod_i p_\theta(y_{t,i} \mid y_{t,<i}, x_t),$$

where $y_{t,i}$ is the i-th token of the target string. This objective allows the model to use the full document context and memory $\mathcal{H}_t$ when predicting the current concept.

Robust Memory Construction via Cross-Validation. During inference, $\mathcal{H}_t$ is filled with the model's own previous predictions, which may be wrong. If the model is trained only with gold concepts in memory, it can over-trust this signal and propagate early errors to later mentions. To reduce this exposure bias, we build the training memory using model predictions instead of gold

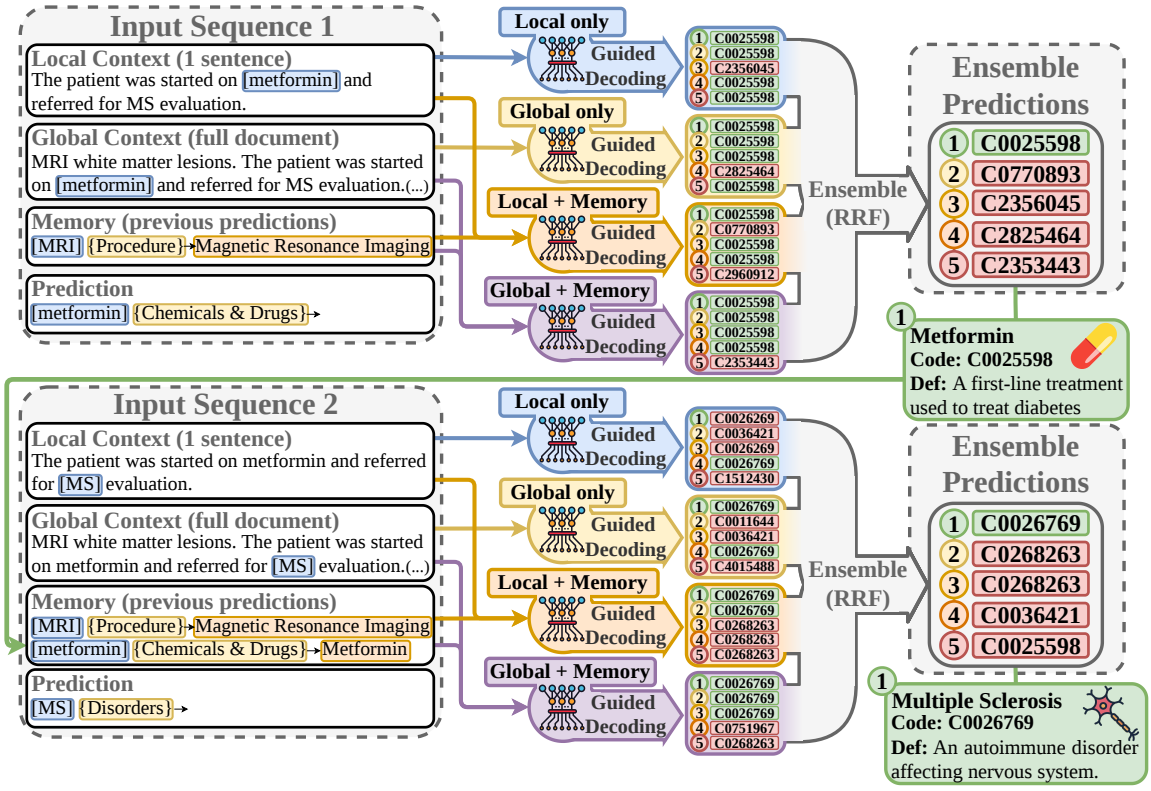

Figure 2: Overview of LongBEL inference. For each mention, LongBEL combines local context, full-document context, and memory of previous predictions to generate multiple ranked candidate lists using semantic-guided constrained decoding. These lists are merged with Reciprocal Rank Fusion (RRF) to produce the final prediction. The predicted concept is then added to memory and used for subsequent mentions.

labels. We use 5-fold cross-validation on the training set: five base models are trained, each on four folds, and each model generates predictions for its held-out fold. These out-of-fold predictions are then used to populate $\mathcal{H}_t$ in the final training data. This exposes LongBEL to realistic inference-time noise and encourages it to balance memory with the current document context. Figure 3 in Appendix A illustrates this process.

3.3 Inference

Sequential Prediction with Dynamic Memory. During inference, as illustrated in Figure 2, LongBEL resolves mentions sequentially as they appear in the document context c. For the first mention m_1, the historical memory $\mathcal{H}_1$ is initialized empty. Once a prediction e_t is generated for mention m_t, the tuple (m_t, e_t) is appended to update the memory state for the next step, $\mathcal{H}_{t+1}$. This dynamic, autoregressive process explicitly propagates prior linking decisions forward, promoting document-level consistency without requiring secondary reranking or graph clustering.

Semantic-Guided Constrained Decoding. To strictly enforce predictions within the target KB, we employ a constrained decoding strategy introduced by GENRE (Cao et al., 2021) and adapted for semantic groups by SynCABEL (Remaki et al., 2026). First, the vocabulary is filtered according to the target entity's given semantic group. Then, autoregressive token generation is constrained dynamically using a prefix trie, restricting the model at each step to only valid continuations.

Ensemble via Reciprocal Rank Fusion. We combine several input configurations using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Each model produces a ranked list of top-K candidate concepts using constrained beam search. For each candidate e, we compute:

$$\mathrm{RRF}(e) = \sum_{m \in M} \frac{1}{k + r_m(e)}, \qquad (1)$$

where $r_m(e)$ is the rank of e in model m, and unranked candidates receive a score of zero. Candidates are sorted by their RRF score, and the top candidate is selected as the final prediction.

4 Experiments

4.1 Datasets

We evaluate LongBEL on five datasets drawn from three annotated corpora spanning three languages: **MedMentions-ST21pv (MM-ST21pv)** (Mohan and Li, 2019), which consists of English PubMed abstracts; **QUAERO-EMEA** (Névéol et al., 2014), comprising French regulatory documents; and the **Spanish Clinical Case Corpus (SPACCC)**. SPACCC is a single shared collection of 1,000 real-world clinical reports that provides three independent annotations, which we evaluate as separate datasets: **SympTEMIST** for symptoms (Miranda-Escalada et al., 2022), **DisTEMIST** for diseases (Lima-López et al., 2023a), and **MedProcNER** for procedures (Lima-López et al., 2023b).

These corpora were selected because they represent large human-annotated BEL benchmarks in their respective languages. We evaluate the EMEA subset of QUAERO and omit QUAERO-MEDLINE, as the latter contains only short title-length texts and is not suitable for document-level context evaluation. Table 1 summarizes the main dataset statistics. The datasets differ strongly in concept recurrence: MM-ST21pv and QUAERO-EMEA have high recurrence rates (51.5% and 58.8% of mentions correspond to recurring concepts), while the SPACCC clinical datasets have much lower rates (11.2%, 15.4%, and 16.3%). This variation allows us to evaluate LongBEL under different levels of document-level redundancy.

	MM-ST21pv	QUAERO-EMEA	Symp-TEMIST	Dis-TEMIST	MedProc-NER
# Docs	4,392	26	1,000	1,000	1,000
Avg. Len.	1,581	5,062	2,337	2,337	2,337
Language	En.	Fr.	Es.	Es.	Es.
Clinical	✗	✗	✓	✓	✓
# Mentions	203,185	7,127	11,683	7,041	8,379
# Concepts	25,419	1,226	4,161	3,015	2,351
# Sem. Grps	14	10	1	1	1
% First	48.5%	41.2%	88.8%	84.6%	83.7%
% Recurring	51.5%	58.8%	11.2%	15.4%	16.3%

Table 1: Summary statistics of the evaluation datasets. # Sem. Grps: number of semantic groups; Avg. Len.: average document length in characters; Clinical: whether the dataset consists of clinical reports; % First: percentage of mentions whose concept appears for the first time in a document; % Recurring: percentage of mentions whose concept has already appeared at least once in the same document.

4.2 Implementation Details

We implement LongBEL with PyTorch and Hugging Face Transformers, using one NVIDIA H100 80GB GPU per run. We use UMLS for MM-ST21pv and QUAERO, and SNOMED CT for SPACCC. Appendix B provides KB statistics and training/inference hyperparameters.

4.3 Results

4.3.1 Benchmark Performance

Table 2 reports Recall@1 scores with ± half the 95% confidence interval across five BEL benchmarks. We compare LongBEL with state-of-the-art systems from different families, including context-free approaches (SciSpacy, SapBERT, SapBERT-all, CODER-all, BERGAMOT), local-context models such as ArboEL, and generative GENRE-style models. Overall, context-free methods lag behind models that use contextual information, highlighting the importance of context for BEL.

To isolate the effect of long-context modeling, we instantiate GENRE with several backbones (mBART-large, Llama-1B, and Llama-8B). These baselines link each mention at the sentence level, using only the sentence that contains the target mention. In contrast, **LongBEL** uses the same Llama backbones but conditions predictions on the full document and on a robust memory of previous predictions. Therefore, the comparison between GENRE-Llama and LongBEL measures the effect of moving from sentence-level to document-level generative linking.

Long-context modeling is especially beneficial on MM-ST21pv, where **LongBEL-8B** improves over the sentence-level Llama-8B baseline from 75.0 to 79.3. This gain is substantially larger than the confidence interval, indicating a robust improvement from document-level modeling. Gains on the other datasets are more moderate, consistent with their different levels of concept recurrence. We also evaluate an ensemble based on Reciprocal Rank Fusion (RRF), combining four variants: local context, global context only, robust memory only, and global context with robust memory. Across all datasets, the ensemble achieves the best performance, with **LongBEL-8B + Ensemble** obtaining the strongest results overall. This indicates that local context, global context, and memory capture complementary signals.

To ensure a fair comparison, all models are evaluated using the same KBs, with candidate entities

Model	MM-ST21PV (English)	QUAERO-EMEA (French)	Symp-TEMIST (Spanish)	Dis-TEMIST (Spanish)	MedProc-NER (Spanish)
CONTEXT-FREE BEL					
SciSpacy (Neumann et al., 2019)	53.8±1.0	37.1±4.3	9.8±1.3	21.1±1.9	10.3±1.2
SapBERT (Liu et al., 2021a)	65.6±1.0	59.7±3.8	34.2±2.0	38.6±2.6	30.4±2.1
CODER-all (Yuan et al., 2022b)	62.9±1.1	66.9±4.0	42.2±2.2	47.0±2.6	42.7±2.1
SapBERT-all (Liu et al., 2021b)	64.6±1.1	67.9±3.9	49.8±2.4	49.6±2.6	45.1±2.2
BERGAMOT (Sakhovskiy et al., 2024)	60.9±1.1	63.8±4.9	48.0±2.7	48.9±2.4	42.3±2.2
LOCAL-CONTEXT BEL					
ArboEL (Agarwal et al., 2022)	76.9±0.9	63.0±3.9	55.4±2.5	54.7±2.6	59.7±2.6
GENRE (Cao et al., 2021)					
mBART-large (Tang et al., 2020)	69.6±1.0	69.3±5.4	59.8±2.7	58.7±2.7	66.0±2.3
Llama-1B (Grattafiori et al., 2024)	73.1±1.0	75.1±3.6	60.5±2.4	62.5±2.3	67.4±2.1
Llama-8B (Grattafiori et al., 2024)	75.0±0.9	73.8±4.0	61.7±2.5	63.2±2.5	68.3±2.2
GLOBAL-CONTEXT BEL (OUR METHODS)					
LongBEL-1B	77.6±0.9	74.5±3.7	59.8±2.5	61.9±2.4	66.6±2.1
+ *Ensemble*	78.6±0.8	77.2±3.0	61.8±2.5	64.3±2.2	69.0±2.0
LongBEL-8B	79.3±0.8	75.4±3.4	62.0±2.6	63.6±2.1	69.0±2.1
+ *Ensemble*	**80.0**±0.8	**77.6**±3.0	**63.3**±2.5	**65.8**±2.2	**71.0**±2.0

Table 2: Entity linking performance (Recall@1) on biomedical benchmarks. **LongBEL** denotes our primary architecture utilizing both global document context and robust memory. + *Ensemble* denotes the Reciprocal Rank Fusion. Best results are in **bold**, second-best are underlined. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level.

restricted to the semantic group of each mention. This enforces a shared candidate space for all neural models. The only exception is SciSpacy, whose rule-based design relies on the full KB and does not support such filtering.

4.3.2 Ablation Study

To better understand the role of global context (Ctx) and robust memory (Mem), we conduct an ablation study with both LongBEL-1B and LongBEL-8B, as reported in Table 3. On MM-ST21pv, both global context and memory help improve performance over the baseline. The best single-model results are obtained when the two components are used together, suggesting that both sources of document-level information are useful on this dataset. Global context allows the model to use evidence from the whole document, while memory gives it access to previous linking decisions that may help maintain consistency across mentions.

On the remaining datasets, improvements are more moderate and less consistent, with differences often within confidence intervals. This was already observed in the main benchmark results and sug-

Ctx	Mem.	MM	EMEA	Symp-TEMIST	Dis-TEMIST	MedProc-NER
Llama-1B						
✗	✗	73.1±1.0	75.1±3.6	60.5±2.4	62.5±2.3	67.4±2.1
✓	✗	76.5±0.9	76.0±2.8	60.3±2.6	62.0±2.3	67.5±2.1
✗	✓	77.4±0.8	74.0±3.2	60.1±2.7	62.0±2.3	67.2±2.1
✓	✓	77.6±0.9	74.5±3.7	59.8±2.5	61.9±2.4	66.6±2.1
Ensemble		**78.6**±0.8	**77.2**±3.0	**61.8**±2.5	**64.3**±2.2	**69.0**±2.0
Llama-8B						
✗	✗	75.0±0.9	73.8±4.0	61.7±2.5	63.2±2.5	68.3±2.2
✓	✗	78.6±0.9	73.9±3.1	61.3±2.6	63.3±2.3	69.1±2.2
✗	✓	79.0±0.9	75.3±3.0	61.8±2.4	62.5±2.4	68.2±2.3
✓	✓	79.3±0.8	75.4±3.4	62.0±2.6	63.6±2.1	69.0±2.1
Ensemble		**80.0**±0.8	**77.6**±3.0	**63.3**±2.5	**65.8**±2.2	**71.0**±2.0

Table 3: Ablation study (recall) evaluating LongBEL-1B and LongBEL-8B. 'Ctx' denotes the inclusion of global context, and 'Mem.' indicates whether memory is enabled (✓) or disabled (✗). Best results are in **bold**, second-best are underlined. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level. The ensemble rows fuse all components via RRF.

gests that the usefulness of document-level information varies across benchmarks. We also observe that larger models (8B) integrate global context and memory more reliably, whereas smaller models (1B) exhibit less stable behavior when combining these components.

Finally, in the main Recall@1 comparison, the ensemble obtains the best result on each dataset. This suggests that local context, global context, and memory capture partly different useful signals, and that combining these signals through ensembling leads to the most robust overall performance.

5 Discussion

5.1 Impact of Concept Redundancy

The improvements brought by LongBEL mainly come from its ability to enforce consistency across concepts that recur within a document. In Table 4, we use *recurring concepts* to refer to concepts that have already appeared earlier in the same document. Gains are mainly concentrated on these recurring concepts, while performance on first occurrences stays similar or slightly decreases. For example, on MM-ST21pv, performance on recurring concepts increases from 70.1 to 78.4 for the 8B model, while performance on first occurrences remains nearly unchanged. This suggests that LongBEL does not mainly improve local disambiguation, which is already well handled by sentence-level models. Instead, it helps by sharing information across mentions of the same concept. This is especially useful for repeated concepts, where earlier mentions can disambiguate later underspecified forms.

This explains the dataset-dependent trends observed in the results. The effectiveness of LongBEL depends on how often concepts recur within a document. Datasets with a high proportion of recurring concepts, such as MM-ST21pv and QUAERO-EMEA (51.5% and 58.8%, respectively), provide more opportunities for LongBEL to exploit consistency signals. In contrast, the SPACCC clinical datasets contain far fewer recurring concepts (around 10–16%), which limits the model's impact. Nevertheless, QUAERO-EMEA shows only moderate gains despite its high redundancy rate, suggesting that recurrence alone is not sufficient. The fine-grained semantic group analysis in Appendix C.1 provides one possible explanation: although a large proportion of QUAERO-EMEA mentions correspond to recurring concepts, the absolute number of examples per semantic group is often small, leading to less stable gains. This suggests that LongBEL needs recurring concepts, but also enough examples within each semantic group to exploit consistency signals reliably.

	Llama-1B		**Llama-8B**	
Subset	Baseline	LongBEL	Baseline	LongBEL
MEDMENTIONS-ST21PV				
Overall	73.1±1.0	77.6±0.9	75.0±0.9	79.3±0.8
First	79.1±0.7	79.1±0.7	80.2±0.7	80.3±0.7
Recurring	67.5±1.4	76.1±1.4	70.1±1.5	78.4±1.3
QUAERO-EMEA				
Overall	75.1±3.6	74.5±3.7	73.8±4.0	75.4±3.4
First	75.0±3.4	72.6±3.6	73.6±3.2	74.2±3.5
Recurring	75.2±5.4	75.8±5.2	74.0±5.6	76.3±4.2
SYMPTEMIST				
Overall	60.5±2.4	59.8±2.5	61.7±2.5	62.0±2.6
First	58.0±2.5	57.6±2.5	59.6±2.7	59.8±2.5
Recurring	79.0±5.0	75.6±5.3	76.5±5.3	77.9±5.1
DISTEMIST				
Overall	62.5±2.3	61.9±2.4	63.2±2.5	63.6±2.1
First	61.6±2.5	60.7±2.4	62.2±2.7	61.8±2.2
Recurring	67.4±5.3	68.6±5.5	68.9±5.3	74.3±4.8
MEDPROCNER				
Overall	67.4±2.1	66.6±2.1	68.3±2.2	69.0±2.1
First	66.2±2.1	65.1±2.2	67.1±2.1	67.4±2.3
Recurring	73.7±4.5	74.8±4.2	75.3±4.3	77.6±3.8

Table 4: Recall comparison across datasets, broken down by concept frequency within documents for both the 1B and 8B models. "Baseline" corresponds to the underlying Llama models applied at the sentence level without cross-mention memory. "LongBEL" extends this model with global document context and robust prediction memory. "First" denotes mentions whose concept appears for the first time in the document, while "Recurring" denotes mentions whose concept has already appeared earlier in the document. 95% confidence intervals were estimated via the empirical bootstrap method at the document level.

Overall, these results show that LongBEL acts as a document-level consistency mechanism. Its effectiveness depends on whether concepts recur within a document and is influenced by both concept redundancy and sample size, rather than by improvements on isolated mentions. Appendix D provides detailed subset results, including recurrence rank, lexical-match categories, mention length, and composite concepts in the SPACCC datasets. Appendix E provides qualitative examples illustrating when document-level evidence helps recover the correct concept, when memory propagates errors, and when both systems behave similarly.

5.2 Robust Memory Mitigates Cascading Errors

LongBEL relies on previous predictions, which creates a risk of error propagation. If the model is trained with gold memory, it can learn to over-trust this signal: an early incorrect prediction may then be copied to later mentions of the same concept. To reduce this mismatch between training and inference, LongBEL uses cross-validated predictions during training, as described in Section 3.2.

We evaluate this effect using the *Copy Wrong Memory Error Rate*, defined in Appendix G as the percentage of recurring concepts for which the model repeats an initial incorrect prediction instead of recovering the correct label. The appendix analysis shows that robust memory reduces this error across model sizes and datasets, with reductions of up to 20.7 percentage points on MM-ST21pv, 4.5 points on QUAERO-EMEA, and 10.5 points on SPACCC. This suggests that exposing the model to noisy memory during training helps it use historical context more cautiously, rather than blindly copying previous predictions.

We further illustrate this behavior with the saliency maps in Figure 6, provided in Appendix F. The ND and PET examples show how prediction memory can help LongBEL use earlier definitions and clinical context to resolve ambiguous abbreviations, while the IFN-γ example shows how an incorrect previous normalization can lead to a cascading error.

5.3 Efficiency Trade-offs

BEL is often used to structure clinical text for applications such as cohort discovery, automated coding, decision support, and trial recruitment (Remaki et al., 2025; French and McInnes, 2023), making deployment cost important. We therefore report inference speed and memory footprint in Table 5.

The model families show different trade-offs. Bi-encoders such as SapBERT-all provide high throughput (575.5 mentions/s), but require a large dense candidate index (20.1 GB). ArboEL has a smaller model footprint (1.2 GB), but its cross-encoder reranking lowers throughput (38.9 mentions/s). Autoregressive models such as GENRE and LongBEL instead use a compact candidate trie (5.4 GB), whose size is independent of the backbone.

LongBEL has the same model memory footprint as sentence-level GENRE, since it adds no trainable parameters. Its main cost is computational: longer contexts and memory updates reduce throughput, e.g., from 69.6 mentions/s for GENRE-1B to 48.5 mentions/s for LongBEL-1B. This makes LongBEL most suitable when document-level consistency is more important than maximum throughput.

	Model (GB)	Cand. (GB)	Speed (/s)
SapBERT-all	2.1	20.1	**575.5**
ArboEL	**1.2**	7.1	38.9
GENRE (mBART)	2.3	**5.4**	102.0
GENRE (Llama-1B)	2.4	**5.4**	69.6
GENRE (Llama-8B)	28.6	**5.4**	38.2
LongBEL-1B	2.4	**5.4**	48.5
LongBEL-8B	28.6	**5.4**	15.2

Table 5: Inference speed and memory footprint. Model: model size; Cand.: candidate memory size.

6 Conclusion and Future Work

We presented **LongBEL**, a document-level generative framework for BEL. LongBEL combines full-document context with a memory of previous predictions, allowing the model to improve consistency across concepts that recur within a document. Across five biomedical benchmarks, LongBEL improves over sentence-level generative baselines, with the strongest gains on datasets where concept recurrence is frequent. Our analyses show that these improvements mainly come from better document-level consistency rather than improved isolated mention disambiguation.

Future work will explore two main directions. First, the ensemble results show that local context, global context, and memory provide complementary signals, but this comes with higher inference cost. A practical direction is to apply these components selectively, for example using the full ensemble only for ambiguous mentions or for concepts that recur within a document. Second, our analysis suggests that small sample size can limit the benefit of document-level consistency. Generating synthetic annotated clinical reports could provide more examples of recurring concepts, make LongBEL easier to train in low-resource settings, and help the model learn to use memory more reliably.

7 Limitations

LongBEL has several limitations. First, our experiments assume that mention spans and semantic groups are given, following the setup of the evaluated BEL benchmarks. This allows us to focus on the linking problem itself, but it does not cover the full end-to-end setting. In real applications, LongBEL would need to be preceded by a named entity recognition (NER) system to detect biomedical mentions, and possibly by a classifier to assign semantic groups. Errors from these upstream components could propagate to the linking stage and affect the final predictions.

Second, the benefit of LongBEL depends on the structure of the dataset. Our analysis shows that the model is most useful when concepts recur within documents, because memory can then provide useful consistency signals. When most concepts appear only once, memory has little information to exploit, and gains over sentence-level models are smaller. However, recurrence alone is not sufficient. The results on QUAERO-EMEA suggest that even when a high proportion of mentions correspond to recurring concepts, gains may remain moderate if the absolute number of examples per semantic group is small. This means that LongBEL is not expected to provide the same improvement on all biomedical corpora, and is most effective when both recurrence and sample size are sufficient.

Third, LongBEL relies on sequential memory. During inference, each prediction is added to memory and can influence later predictions in the same document. Robust memory training reduces the risk of blindly copying wrong predictions, but it does not fully remove error propagation. If an early prediction is incorrect and later mentions are ambiguous, the model may still reuse the wrong information.

Finally, LongBEL has a higher inference cost than sentence-level baselines. It processes longer document contexts, updates memory after each prediction, and, in the ensemble setting, requires multiple forward passes per mention. This makes LongBEL less suitable for applications where maximum throughput is the main constraint. It is better suited for settings where document-level consistency and linking accuracy are more important than raw speed.

8 Ethical Considerations

This work uses publicly available biomedical entity linking benchmarks. We do not collect new patient data or access private electronic health records. Clinical datasets used in this study are released as de-identified research benchmarks by their original providers.

LongBEL may still produce incorrect normalizations, which could affect downstream clinical applications such as coding, cohort selection, or decision support. It should therefore not be used for clinical decisions without validation, monitoring, and appropriate human oversight.

Our method requires more computation than sentence-level baselines because it uses document-level context and robust memory construction. Future work should consider the trade-off between performance, inference cost, and environmental impact.

All datasets and resources were used according to their licenses and access conditions, including MedMentions, QUAERO, SPACCC, UMLS, and SNOMED CT. The authors used GitHub Copilot for code completion; all generated code was reviewed by the authors.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States.

Yihao Ai, Zhiyuan Ning, Weiwei Dai, Pengfei Wang, Yi Du, Wenjuan Cui, Kunpeng Liu, and Yuanchun Zhou. 2025. Distilling Closed-Source LLM's Knowledge for Locally Stable and Economic Biomedical Entity Linking. In *Advanced Intelligent Computing Technology and Applications*, Singapore.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270.

Florian Borchert, Ignacio Llorca, and Matthieu-P Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database*, 2024:baae067.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*.

Haihua Chen, Yuhan Zhou, Ruochi Li, Aryan Murthy Illa, Ana Cleveland, and Junhua Ding. 2026. A comprehensive survey on medical concept normalization: Datasets, techniques, applications, and future directions. *Journal of Biomedical Informatics*, 178:105005.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, New York, NY, USA.

Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. 2025. EntGPT: Entity Linking with Generative Large Language Models. *arXiv preprint*. ArXiv:2402.06738 [cs].

Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024. ChatEL: Entity Linking with Chatbots. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia.

Kevin Donnelly. 2006. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. In *Medical and Care Compunetics 3*, pages 279–290. IOS Press.

Evan French and Bridget T. McInnes. 2023. An overview of Biomedical Entity Linking throughout the years. *Journal of biomedical informatics*, 137:104252.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. 2025. Learning from Negative Samples in Biomedical Generative Entity Linking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10714–10730, Vienna, Austria. Association for Computational Linguistics.

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526.

Salvador Lima-López, Eulalia Farré-Maduell, Luis Gasco Sanchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023a. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023b. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. *Working Notes of CLEF*.

Zhenxi Lin, Ziheng Zhang, Jian Wu, Yefeng Zheng, and Xian Wu. 2025. Guiding Large Language Models for Biomedical Entity Linking via Restrictive and Contrastive Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23745–23759, Suzhou, China. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *arXiv preprint*. ArXiv:1902.09476 [cs].

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Adam Remaki, Christel Gérardin, Eulàlia Farré-Maduell, Martin Krallinger, and Xavier Tannier. 2026. SynCABEL: Synthetic Contextualized Augmentation for Biomedical Entity Linking. *arXiv preprint*. ArXiv:2601.19667 [cs].

Adam Remaki, Jacques Ung, Pierre Pages, Perceval Wajsburt, Elise Liu, Guillaume Faure, Thomas Petit-Jean, Xavier Tannier, and Christel Gérardin. 2025. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study. *JMIR Medical Informatics*, 13(1):e68704. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643, Mexico City, Mexico. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Ariel S. Schwartz and Marti A. Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003*, pages 451–462. WORLD SCIENTIFIC.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389, Bangkok, Thailand. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv preprint*. ArXiv:2008.00401 [cs].

Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. Contextual Augmentation for Entity Linking using Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8535–8545, Abu Dhabi, UAE. Association for Computational Linguistics.

Fang Wang, Zhengwei Tao, Ming Wang, Minghao Hu, and Xiaoying Bai. 2025. AELC: Adaptive Entity Linking with LLM-Driven Contextualization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4313–4327, Suzhou, China. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Yuzhang Xie, Jiaying Lu, Joyce Ho, Fadi Nahab, Xiao Hu, and Carl Yang. 2024. PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 2589–2593, New York, NY, USA. Association for Computing Machinery.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2025. LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 3550–3559, New York, NY, USA. Association for Computing Machinery.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022a. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Lixing Zhu, Jun Wang, and Yulan He. 2025. Llm-Link: Dual LLMs for Dynamic Entity Linking on Long Narratives with Collaborative Memorisation and Prompt Optimisation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11334–11347, Abu Dhabi, UAE. Association for Computational Linguistics.

A Robust Memory Illustration

Figure 3 illustrates how cross-validated predictions are used to construct realistic memory for robust LongBEL training.

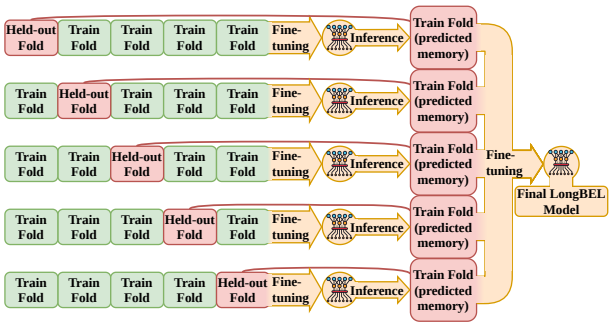

Figure 3: Illustration of robust memory construction via cross-validation. The training set is split into five folds. For each fold, a base model is trained on the remaining folds and used to generate predictions on the held-out fold. These held-out predictions are then used to create the memory during final LongBEL training, exposing the model to realistic inference-time memory.

B Extended Implementation Details

Knowledge Bases. For each corpus, we selected a KB consistent with its original annotation setup. For MM-ST21pv, we use UMLS 2017AA, narrowed down to the ST21-pv subset which covers 21 semantic types with synonyms sourced from 18 vocabularies. QUAERO is paired with UMLS 2014AA, from which we retained only the 10 semantic groups relevant to the corpus. For the three SPACCC datasets, candidate concepts were drawn from gazetteers built from the appropriate branches of the Spanish edition of SNOMED-CT (July 31, 2021 release). The statistics of these KB subsets are summarized in Table 6.

	MM-ST21pv	QUAERO	SPACCC
Version	2017AA	2014AA	2021
# Concepts	2,374,392	2,960,365	276,670
# Synonyms	4,905,087	6,970,593	546,814
# Semantic Groups	14	10	3
# Synonyms per Code	2.1	2.4	1.5

Table 6: Summary statistics of the knowledge base subsets used for candidate retrieval across the evaluated datasets.

Training and Inference Setup. We train our models using the `trl` `SFTTrainer` library. All detailed hyperparameters are provided in Tables 7 and 8.

Hyperparameter	Value
Learning Rate	3×10^{-5}
Learning Rate Scheduler	Linear
Warmup Ratio	0.03
Epochs	50
Precision	`BF16`
Attention Implementation	Flash Attention 2
Gradient Accumulation Steps	1
Dataset Max Input Length	$L_{\max}$ tokens
Token Budget per Batch	16,384 tokens
Per-device Batch Size	$\lfloor 16{,}384/L_{\max} \rfloor$

Table 7: Supervised Fine-Tuning (SFT) hyperparameters used across all evaluated models. $L_{\max}$ is the maximum tokenized input length over the training split.

Hyperparameter	Value
Decoding Strategy	Beam Search
Number of Beams	5
Batch Size	16
Precision	`BF16`
RRF k	60

Table 8: Inference hyperparameters used for the generation phase.

C Additional Behavioral Analysis

C.1 Fine-grained Analysis by Semantic Group

We report Δ Recall, defined as the Recall@1 difference between LongBEL and the sentence-level baseline. Figures 4 and 5 show the results for LongBEL-1B and LongBEL-8B by semantic group and dataset. LongBEL improves recall across most semantic groups, showing that its benefits are not limited to one type of entity. On MM-ST21pv, gains are relatively stable across groups, with larger improvements for frequent and ambiguous groups such as *Chemicals*, *Disorders*, and *Procedures*. Smaller semantic groups show more variation, likely because they contain fewer examples. On SPACCC and QUAERO-EMEA, the gains are generally smaller and more variable across groups. This is likely due to both lower concept redundancy and smaller sample sizes in some categories, which make group-level estimates less stable. Overall, these results show that LongBEL provides broad improvements across semantic groups, while the size and stability of the gains depend on both concept redundancy and sample size.

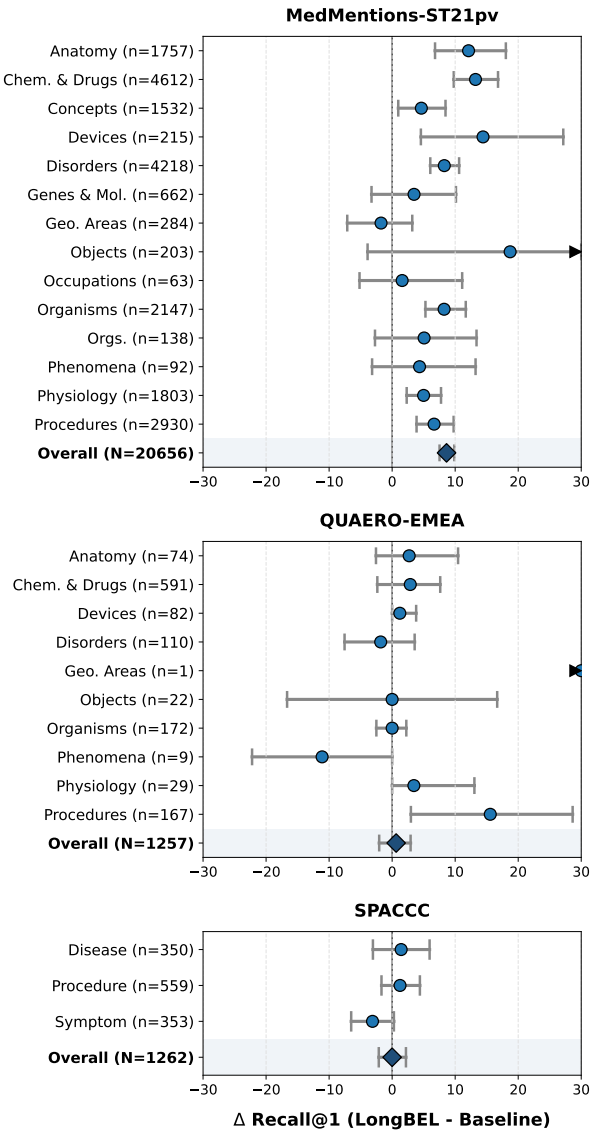

Figure 4: Difference in Recall@1 between LongBEL and the sentence-level baseline, computed on mentions of repeated concepts only, for the LongBEL-1B model. Results are reported per semantic group across datasets.

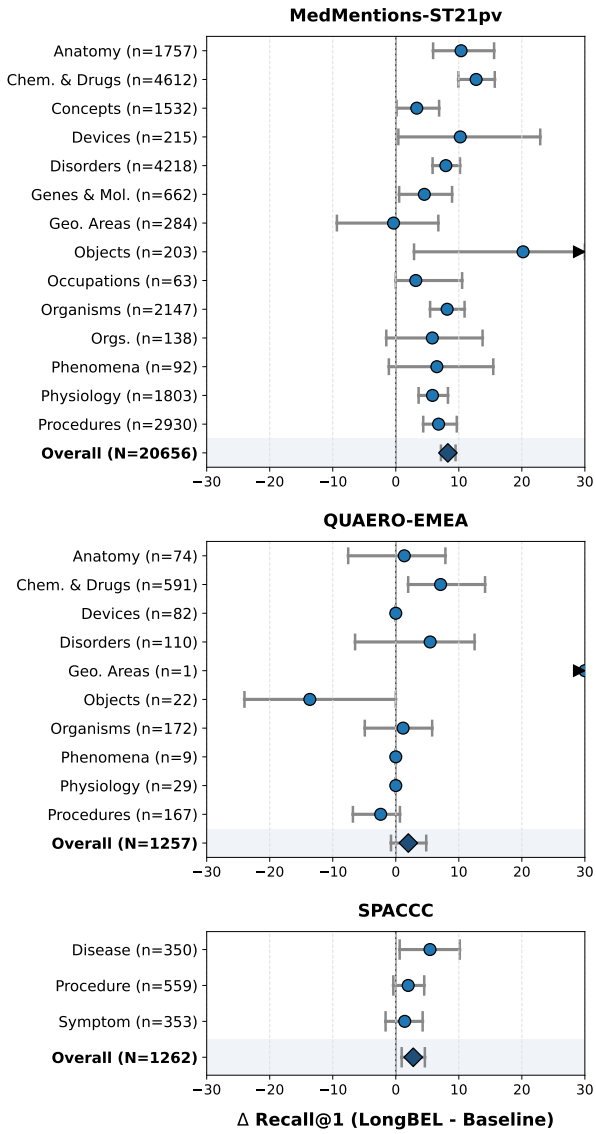

Figure 5: Difference in Recall@1 between LongBEL and the sentence-level baseline, computed on mentions of repeated concepts only, for the LongBEL-8B model. Results are reported per semantic group across datasets.

D Detailed Subset Results

Tables 9–13 report detailed subset results for the sentence-level baselines, LongBEL, and the ensemble. The first two rows report overall Recall@1 and Recall@5. The other rows report Recall@1 on different subsets. The support column gives the percentage and number of examples in each subset. Each score is reported with half of the 95% confidence interval.

The subsets capture simple properties of the mentions. Seen and unseen concepts indicate whether the gold concept appears in the training set. Identical and non-identical mentions indicate whether the mention text exactly matches a synonym of the gold concept in the knowledge base. We also group mentions by length and by document recurrence. First occurrence means that the gold concept has not appeared earlier in the document, while recurring occurrence means that it has already appeared before. For the Spanish SPACCC datasets only, we also separate simple and composite concepts. Simple concepts map to one concept, while composite concepts map to a combination of several concepts.

These tables add detail to the main analysis. MM-ST21pv is much larger than QUAERO-EMEA and SPACCC, so its subset trends are more stable. In MM-ST21pv, LongBEL changes little on first occurrences but improves clearly on recurring men-

tions, with the largest gains on fifth-or-later occurrences. The smaller datasets show less stable trends because some subsets have few examples and large confidence intervals.

The tables also highlight dataset differences. The SPACCC datasets contain many more multi-word mentions than MM-ST21pv and QUAERO-EMEA, which makes the normalization setting harder. They also include composite concepts, which are specific to these datasets and remain difficult for all models.

E Qualitative Examples

Table 14 compares LongBEL-8B with the sentence-level Llama-8B baseline on representative cases: cases where document-level context helps, cases where memory propagates an error, and cases where both systems either fail or succeed.

F Saliency-based Analysis

Figure 6 shows token-level saliency maps for three examples, with all predictions and saliency scores produced by LongBEL-8B. The maps show that the model can use information from document context and previous predictions, which can help but can also propagate earlier errors.

We compute saliency with a simple gradient-based method. After decoding, we keep the top generated sequence. We define the objective as the sum of the log-probabilities of the generated output tokens:

$$S(x, y) = \sum_{t \in y} \log p_\theta(y_t \mid x, y_{<t}),$$

where x is the prompt and y is the generated output. We then compute the gradient of this objective with respect to the input token embeddings. The saliency score of an input token is the ℓ_2 norm of this gradient:

$$a_i = \left\| \frac{\partial S(x, y)}{\partial e_i} \right\|_2,$$

where e_i is the embedding of the i-th input token. We only show scores for prompt tokens. Generated tokens and padding tokens are not shown. For each example, we normalize the scores by the maximum score in that example.

This saliency analysis is a local sensitivity measure. A highlighted token is a token that strongly affects the likelihood of the generated prediction. It is not based on attention weights, and it should not be read as a causal explanation.

G Robustness to Noisy Memory

To evaluate whether a model propagates errors through memory, we define the *Copy Wrong Memory Error Rate* (CWME). CWME measures how often a model repeats an earlier incorrect prediction for the same ground-truth concept.

Let $M_e = (m_1, m_2, \ldots, m_k)$ be the ordered mentions of concept e in a document. Suppose the model first predicts an incorrect concept $e' \neq e$ at mention m_f. All later mentions of e are then *exposed mentions*, since the wrong prediction is present in memory:

$$E_e = \{m_i \in M_e \mid i > f\}.$$

An *echoed error* occurs when the model repeats the same wrong prediction e':

$$R_e = \{m_i \in E_e \mid \hat{e}_i = e'\}.$$

CWME is computed as:

$$\text{CWME} = \frac{\sum_{e \in \mathcal{E}} |R_e|}{\sum_{e \in \mathcal{E}} |E_e|} \times 100,$$

where $\mathcal{E}$ includes concepts with an initial error and at least one later exposed mention. Lower CWME indicates fewer cascading errors. Figures 7 and 8 compare gold-memory training with robust cross-validated memory training using

$$\Delta\text{CWME} = \text{CWME}_{\text{robust memory}} - \text{CWME}_{\text{gold memory}}.$$

Negative values indicate fewer copied errors with robust memory.

Overall, robust memory consistently reduces copied errors. For LongBEL-1B, CWME decreases by 20.7 points on MM-ST21pv, 4.1 points on QUAERO-EMEA, and 6.7 points on SPACCC. For LongBEL-8B, the corresponding reductions are 20.0, 4.5, and 10.5 points, respectively. These results show that training with predicted memory makes the model less likely to repeat earlier mistakes. The reduction is strongest and most stable on MM-ST21pv, where many semantic groups contain enough exposed mentions to provide reliable estimates. In contrast, some groups in QUAERO-EMEA and SPACCC contain very few examples, leading to noisier category-level results and occasional positive deltas. This is consistent with our main analysis: robust memory is most effective when enough recurring concepts are available to learn reliable consistency patterns.

Subset	Support	Baseline-1B	LongBEL-1B	Ensemble-1B	Baseline-8B	LongBEL-8B	Ensemble-8B
Overall performance							
Recall@1 overall	100% (40,143)	$73.1_{\pm 1.0}$	$77.6_{\pm 0.9}$	$78.6_{\pm 0.8}$	$75.0_{\pm 0.9}$	$79.3_{\pm 0.8}$	$80.0_{\pm 0.8}$
Recall@5 overall	100% (40,143)	$81.2_{\pm 0.8}$	$85.8_{\pm 0.7}$	$84.1_{\pm 0.8}$	$83.7_{\pm 0.9}$	$87.2_{\pm 0.8}$	$85.6_{\pm 0.7}$
Concept familiarity							
Seen concepts	79.5% (31,914)	$80.5_{\pm 0.9}$	$82.8_{\pm 0.8}$	$84.0_{\pm 0.8}$	$82.2_{\pm 0.8}$	$84.6_{\pm 0.8}$	$85.3_{\pm 0.7}$
Unseen concepts	20.5% (8,229)	$44.6_{\pm 2.1}$	$57.5_{\pm 2.4}$	$57.7_{\pm 2.4}$	$47.1_{\pm 2.2}$	$59.1_{\pm 2.5}$	$59.6_{\pm 2.5}$
Lexical match							
Identical	5.6% (2,265)	$93.5_{\pm 1.9}$	$95.5_{\pm 1.3}$	$96.0_{\pm 1.3}$	$94.6_{\pm 1.4}$	$94.3_{\pm 1.4}$	$95.5_{\pm 1.5}$
Not identical	94.4% (37,878)	$71.9_{\pm 0.9}$	$76.5_{\pm 0.9}$	$77.6_{\pm 0.9}$	$73.8_{\pm 0.9}$	$78.4_{\pm 0.8}$	$79.1_{\pm 0.9}$
Mention length							
Single word	66.8% (26,833)	$74.9_{\pm 1.1}$	$80.4_{\pm 1.0}$	$81.4_{\pm 1.0}$	$76.6_{\pm 1.1}$	$82.3_{\pm 1.0}$	$82.6_{\pm 0.9}$
Multi-word	33.2% (13,310)	$69.5_{\pm 1.2}$	$72.0_{\pm 1.2}$	$73.0_{\pm 1.1}$	$71.7_{\pm 1.1}$	$73.5_{\pm 1.1}$	$74.7_{\pm 1.1}$
Document recurrence							
First occurrence	48.5% (19,487)	$79.1_{\pm 0.7}$	$79.1_{\pm 0.7}$	$80.6_{\pm 0.6}$	$80.2_{\pm 0.6}$	$80.3_{\pm 0.7}$	$81.7_{\pm 0.6}$
Recurring occurrence	51.5% (20,656)	$67.5_{\pm 1.4}$	$76.1_{\pm 1.4}$	$76.8_{\pm 1.3}$	$70.1_{\pm 1.5}$	$78.4_{\pm 1.3}$	$78.4_{\pm 1.2}$
2nd occurrence	18.5% (7,436)	$75.8_{\pm 1.1}$	$78.6_{\pm 1.0}$	$79.8_{\pm 1.0}$	$77.1_{\pm 1.0}$	$80.4_{\pm 0.9}$	$80.8_{\pm 0.9}$
3rd occurrence	10.4% (4,167)	$69.6_{\pm 1.6}$	$77.1_{\pm 1.4}$	$77.8_{\pm 1.3}$	$72.2_{\pm 1.6}$	$79.2_{\pm 1.4}$	$79.0_{\pm 1.3}$
4th occurrence	6.7% (2,704)	$65.0_{\pm 1.9}$	$75.8_{\pm 1.7}$	$76.3_{\pm 1.7}$	$68.2_{\pm 1.8}$	$77.8_{\pm 1.5}$	$78.1_{\pm 1.5}$
5th+ occurrence	15.8% (6,349)	$57.3_{\pm 2.8}$	$72.7_{\pm 2.5}$	$72.8_{\pm 2.5}$	$61.4_{\pm 2.6}$	$75.9_{\pm 2.4}$	$75.4_{\pm 2.4}$

Table 9: Detailed subset results on MM-ST21pv. The first two rows report overall Recall@1 and Recall@5. All other rows report Recall@1 on specific mention subsets. The support column gives the percentage and number of examples in each subset. Each score is shown with half of the 95% confidence interval. First occurrence refers to mentions whose gold concept has not appeared earlier in the document, while recurring occurrence refers to mentions whose gold concept has already appeared earlier.

Subset	Support	Baseline-1B	LongBEL-1B	Ensemble-1B	Baseline-8B	LongBEL-8B	Ensemble-8B
Overall performance							
Recall@1 overall	100% (2,196)	$75.1_{\pm 3.6}$	$74.5_{\pm 3.7}$	$77.2_{\pm 3.0}$	$73.8_{\pm 4.0}$	$75.4_{\pm 3.4}$	$77.6_{\pm 3.0}$
Recall@5 overall	100% (2,196)	$82.0_{\pm 3.0}$	$82.4_{\pm 2.8}$	$82.8_{\pm 3.2}$	$82.2_{\pm 2.6}$	$82.0_{\pm 2.5}$	$82.8_{\pm 2.7}$
Concept familiarity							
Seen concepts	56.0% (1,229)	$88.7_{\pm 2.7}$	$89.3_{\pm 2.1}$	$91.1_{\pm 1.8}$	$92.1_{\pm 2.1}$	$90.8_{\pm 2.8}$	$92.8_{\pm 1.2}$
Unseen concepts	44.0% (967)	$57.5_{\pm 5.8}$	$55.5_{\pm 5.5}$	$59.2_{\pm 5.6}$	$50.2_{\pm 5.1}$	$55.8_{\pm 5.4}$	$58.0_{\pm 5.6}$
Lexical match							
Identical	6.5% (143)	$93.7_{\pm 4.4}$	$97.9_{\pm 2.5}$	$100.0_{\pm 0.0}$	$93.0_{\pm 4.6}$	$99.3_{\pm 1.0}$	$98.6_{\pm 1.8}$
Not identical	93.5% (2,053)	$73.8_{\pm 3.4}$	$72.8_{\pm 3.9}$	$75.6_{\pm 3.1}$	$72.5_{\pm 4.0}$	$73.7_{\pm 3.6}$	$76.2_{\pm 2.9}$
Mention length							
Single word	77.6% (1,704)	$79.9_{\pm 4.0}$	$80.8_{\pm 4.0}$	$82.1_{\pm 3.5}$	$78.5_{\pm 4.8}$	$80.3_{\pm 4.9}$	$82.0_{\pm 3.7}$
Multi-word	22.4% (492)	$58.5_{\pm 6.4}$	$52.6_{\pm 9.1}$	$60.2_{\pm 8.2}$	$57.7_{\pm 5.7}$	$58.3_{\pm 6.0}$	$62.4_{\pm 5.6}$
Document recurrence							
First occurrence	42.8% (939)	$75.0_{\pm 3.4}$	$72.6_{\pm 3.6}$	$75.6_{\pm 3.5}$	$73.6_{\pm 3.1}$	$74.2_{\pm 3.5}$	$76.3_{\pm 2.8}$
Recurring occurrence	57.2% (1,257)	$75.2_{\pm 5.4}$	$75.8_{\pm 5.1}$	$78.4_{\pm 4.6}$	$74.0_{\pm 5.5}$	$76.3_{\pm 4.9}$	$78.7_{\pm 4.3}$
2nd occurrence	18.9% (416)	$77.4_{\pm 2.7}$	$76.0_{\pm 4.6}$	$79.1_{\pm 3.3}$	$76.2_{\pm 4.4}$	$77.6_{\pm 3.2}$	$79.3_{\pm 3.2}$
3rd occurrence	9.6% (210)	$72.4_{\pm 5.7}$	$71.4_{\pm 5.4}$	$75.7_{\pm 4.1}$	$72.9_{\pm 4.4}$	$74.8_{\pm 3.5}$	$75.2_{\pm 4.2}$
4th occurrence	5.5% (120)	$77.5_{\pm 5.3}$	$76.7_{\pm 5.0}$	$78.3_{\pm 4.8}$	$75.8_{\pm 6.0}$	$76.7_{\pm 5.5}$	$79.2_{\pm 4.4}$
5th+ occurrence	23.3% (511)	$74.0_{\pm 9.6}$	$77.3_{\pm 9.2}$	$78.9_{\pm 8.6}$	$72.2_{\pm 10.1}$	$75.7_{\pm 9.7}$	$79.5_{\pm 8.2}$

Table 10: Detailed subset results on QUAERO-EMEA. The first two rows report overall Recall@1 and Recall@5. All other rows report Recall@1 on specific mention subsets. The support column gives the percentage and number of examples in each subset. Each score is shown with half of the 95% confidence interval. First occurrence refers to mentions whose gold concept has not appeared earlier in the document, while recurring occurrence refers to mentions whose gold concept has already appeared earlier.

Subset	Support	Baseline-1B	LongBEL-1B	Ensemble-1B	Baseline-8B	LongBEL-8B	Ensemble-8B
Overall performance							
Recall@1 overall	100% (2,898)	$60.5_{\pm 2.4}$	$59.8_{\pm 2.5}$	$61.8_{\pm 2.5}$	$61.7_{\pm 2.5}$	$62.0_{\pm 2.6}$	$63.3_{\pm 2.5}$
Recall@5 overall	100% (2,898)	$68.2_{\pm 2.4}$	$66.1_{\pm 2.4}$	$70.5_{\pm 2.2}$	$69.6_{\pm 2.4}$	$68.8_{\pm 2.4}$	$70.3_{\pm 2.3}$
Concept familiarity							
Seen concepts	71.2% (2,064)	$78.2_{\pm 2.1}$	$77.4_{\pm 2.2}$	$79.2_{\pm 2.0}$	$79.1_{\pm 2.1}$	$79.3_{\pm 2.1}$	$80.6_{\pm 2.0}$
Unseen concepts	28.8% (834)	$16.7_{\pm 2.8}$	$16.2_{\pm 2.4}$	$18.8_{\pm 2.8}$	$18.6_{\pm 2.7}$	$19.3_{\pm 2.8}$	$20.3_{\pm 2.8}$
Lexical match							
Identical	21.4% (621)	$96.3_{\pm 1.4}$	$96.1_{\pm 1.6}$	$97.7_{\pm 1.4}$	$96.8_{\pm 1.5}$	$97.4_{\pm 1.2}$	$96.8_{\pm 1.5}$
Not identical	78.6% (2,277)	$50.8_{\pm 2.6}$	$49.8_{\pm 2.8}$	$52.0_{\pm 2.6}$	$52.1_{\pm 2.9}$	$52.3_{\pm 2.8}$	$54.1_{\pm 2.8}$
Mention length							
Single word	27.6% (800)	$93.4_{\pm 1.8}$	$92.6_{\pm 2.4}$	$93.5_{\pm 2.1}$	$93.8_{\pm 2.0}$	$93.9_{\pm 1.9}$	$93.8_{\pm 2.1}$
Multi-word	72.4% (2,098)	$48.0_{\pm 2.8}$	$47.2_{\pm 2.9}$	$49.7_{\pm 2.8}$	$49.5_{\pm 2.8}$	$49.9_{\pm 2.9}$	$51.6_{\pm 2.9}$
Document recurrence							
First occurrence	87.8% (2,545)	$58.0_{\pm 2.5}$	$57.6_{\pm 2.6}$	$59.4_{\pm 2.5}$	$59.6_{\pm 2.7}$	$59.8_{\pm 2.6}$	$60.9_{\pm 2.6}$
Recurring occurrence	12.2% (353)	$79.0_{\pm 5.0}$	$75.6_{\pm 5.3}$	$79.3_{\pm 4.8}$	$76.5_{\pm 5.2}$	$77.9_{\pm 5.1}$	$80.2_{\pm 5.1}$
2nd occurrence	8.6% (249)	$74.7_{\pm 5.7}$	$70.3_{\pm 5.6}$	$74.3_{\pm 5.7}$	$71.5_{\pm 5.6}$	$71.9_{\pm 6.0}$	$74.7_{\pm 5.9}$
3rd occurrence	2.1% (62)	$87.1_{\pm 8.7}$	$85.5_{\pm 8.9}$	$88.7_{\pm 8.3}$	$83.9_{\pm 10.4}$	$90.3_{\pm 7.9}$	$90.3_{\pm 7.9}$
4th occurrence	0.7% (21)	$90.5_{\pm 11.9}$	$90.5_{\pm 11.9}$	$90.5_{\pm 11.9}$	$90.5_{\pm 11.9}$	$90.5_{\pm 11.9}$	$95.2_{\pm 8.4}$
5th+ occurrence	0.7% (21)	$95.2_{\pm 8.8}$	$95.2_{\pm 8.4}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Concept structure							
Simple concepts	94.2% (2,730)	$62.6_{\pm 2.5}$	$61.8_{\pm 2.7}$	$64.1_{\pm 2.4}$	$63.7_{\pm 2.6}$	$64.5_{\pm 2.5}$	$65.1_{\pm 2.7}$
Composite concepts	5.8% (168)	$26.8_{\pm 7.4}$	$26.8_{\pm 7.1}$	$24.4_{\pm 6.7}$	$29.8_{\pm 7.0}$	$21.4_{\pm 5.9}$	$32.7_{\pm 7.1}$

Table 11: Detailed subset results on SympTEMIST. The first two rows report overall Recall@1 and Recall@5. All other rows report Recall@1 on specific mention subsets. The support column gives the percentage and number of examples in each subset. Each score is shown with half of the 95% confidence interval. First occurrence refers to mentions whose gold concept has not appeared earlier in the document, while recurring occurrence refers to mentions whose gold concept has already appeared earlier. Simple and composite concept rows are reported only for the Spanish SPACCC datasets.

Subset	Support	Baseline-1B	LongBEL-1B	Ensemble-1B	Baseline-8B	LongBEL-8B	Ensemble-8B
Overall performance							
Recall@1 overall	100% (2,341)	$62.5_{\pm 2.3}$	$61.9_{\pm 2.4}$	$64.3_{\pm 2.2}$	$63.2_{\pm 2.5}$	$63.6_{\pm 2.1}$	$65.8_{\pm 2.2}$
Recall@5 overall	100% (2,341)	$69.5_{\pm 2.1}$	$68.9_{\pm 2.1}$	$73.1_{\pm 2.1}$	$71.1_{\pm 2.2}$	$69.9_{\pm 2.1}$	$73.2_{\pm 2.1}$
Concept familiarity							
Seen concepts	61.7% (1,445)	$83.0_{\pm 2.2}$	$80.7_{\pm 2.4}$	$83.9_{\pm 2.1}$	$83.8_{\pm 2.2}$	$84.6_{\pm 1.9}$	$86.4_{\pm 1.8}$
Unseen concepts	38.3% (896)	$29.2_{\pm 3.2}$	$31.5_{\pm 3.3}$	$32.8_{\pm 3.4}$	$30.0_{\pm 3.5}$	$29.9_{\pm 3.4}$	$32.5_{\pm 3.7}$
Lexical match							
Identical	24.9% (583)	$94.2_{\pm 1.9}$	$93.1_{\pm 2.0}$	$95.2_{\pm 1.9}$	$93.3_{\pm 2.2}$	$93.5_{\pm 2.0}$	$95.0_{\pm 1.8}$
Not identical	75.1% (1,758)	$51.9_{\pm 2.6}$	$51.5_{\pm 2.7}$	$54.1_{\pm 2.5}$	$53.2_{\pm 2.8}$	$53.8_{\pm 2.5}$	$56.1_{\pm 2.6}$
Mention length							
Single word	28.3% (663)	$86.7_{\pm 3.4}$	$85.2_{\pm 3.5}$	$87.6_{\pm 3.4}$	$86.0_{\pm 3.5}$	$87.6_{\pm 3.1}$	$88.2_{\pm 3.1}$
Multi-word	71.7% (1,678)	$52.9_{\pm 2.7}$	$52.6_{\pm 2.8}$	$55.1_{\pm 2.6}$	$54.2_{\pm 2.9}$	$54.2_{\pm 2.6}$	$56.9_{\pm 2.8}$
Document recurrence							
First occurrence	85.0% (1,991)	$61.6_{\pm 2.4}$	$60.7_{\pm 2.5}$	$62.9_{\pm 2.3}$	$62.2_{\pm 2.7}$	$61.8_{\pm 2.1}$	$63.9_{\pm 2.4}$
Recurring occurrence	15.0% (350)	$67.4_{\pm 5.3}$	$68.6_{\pm 5.6}$	$72.3_{\pm 5.1}$	$68.9_{\pm 5.3}$	$74.3_{\pm 4.8}$	$76.3_{\pm 4.6}$
2nd occurrence	11.5% (269)	$66.5_{\pm 5.6}$	$67.3_{\pm 5.6}$	$71.0_{\pm 5.0}$	$68.8_{\pm 5.3}$	$72.9_{\pm 5.2}$	$75.8_{\pm 5.1}$
3rd occurrence	2.4% (56)	$67.9_{\pm 10.8}$	$71.4_{\pm 10.5}$	$76.8_{\pm 9.9}$	$67.9_{\pm 10.8}$	$75.0_{\pm 9.6}$	$75.0_{\pm 9.6}$
4th occurrence	0.6% (13)	$69.2_{\pm 23.9}$	$69.2_{\pm 26.7}$	$69.2_{\pm 23.9}$	$76.9_{\pm 25.0}$	$76.9_{\pm 25.0}$	$76.9_{\pm 25.0}$
5th+ occurrence	0.5% (12)	$83.3_{\pm 17.4}$	$83.3_{\pm 17.4}$	$83.3_{\pm 17.4}$	$66.7_{\pm 40.0}$	$100.0_{\pm 0.0}$	$91.7_{\pm 8.7}$
Concept structure							
Simple concepts	96.4% (2,257)	$64.3_{\pm 2.2}$	$63.7_{\pm 2.4}$	$66.3_{\pm 2.2}$	$65.0_{\pm 2.4}$	$65.6_{\pm 2.0}$	$67.7_{\pm 2.1}$
Composite concepts	3.6% (84)	$13.1_{\pm 7.5}$	$13.1_{\pm 7.1}$	$11.9_{\pm 6.8}$	$16.7_{\pm 8.4}$	$11.9_{\pm 6.5}$	$15.5_{\pm 7.5}$

Table 12: Detailed subset results on DisTEMIST. The first two rows report overall Recall@1 and Recall@5. All other rows report Recall@1 on specific mention subsets. The support column gives the percentage and number of examples in each subset. Each score is shown with half of the 95% confidence interval. First occurrence refers to mentions whose gold concept has not appeared earlier in the document, while recurring occurrence refers to mentions whose gold concept has already appeared earlier. Simple and composite concept rows are reported only for the Spanish SPACCC datasets.

Subset	Support	Baseline-1B	LongBEL-1B	Ensemble-1B	Baseline-8B	LongBEL-8B	Ensemble-8B
Overall performance							
Recall@1 overall	100% (3,582)	$67.4_{\pm 2.1}$	$66.6_{\pm 2.1}$	$69.0_{\pm 2.0}$	$68.3_{\pm 2.2}$	$69.0_{\pm 2.1}$	$71.0_{\pm 2.0}$
Recall@5 overall	100% (3,582)	$72.9_{\pm 1.9}$	$71.7_{\pm 2.1}$	$76.1_{\pm 1.9}$	$74.8_{\pm 1.8}$	$74.4_{\pm 1.9}$	$75.7_{\pm 2.0}$
Concept familiarity							
Seen concepts	73.6% (2,636)	$84.3_{\pm 1.8}$	$83.5_{\pm 1.8}$	$85.9_{\pm 1.7}$	$86.2_{\pm 1.6}$	$86.4_{\pm 1.6}$	$87.6_{\pm 1.6}$
Unseen concepts	26.4% (946)	$20.1_{\pm 2.8}$	$19.7_{\pm 2.8}$	$21.8_{\pm 3.0}$	$18.6_{\pm 2.9}$	$20.6_{\pm 2.9}$	$24.9_{\pm 3.2}$
Lexical match							
Identical	18.5% (664)	$97.4_{\pm 1.4}$	$94.9_{\pm 2.0}$	$97.9_{\pm 1.6}$	$97.0_{\pm 1.5}$	$96.5_{\pm 1.8}$	$98.8_{\pm 1.0}$
Not identical	81.5% (2,918)	$60.5_{\pm 2.3}$	$60.2_{\pm 2.4}$	$62.4_{\pm 2.3}$	$61.9_{\pm 2.4}$	$62.6_{\pm 2.4}$	$64.7_{\pm 2.3}$
Mention length							
Single word	33.6% (1,202)	$84.7_{\pm 2.8}$	$83.9_{\pm 2.9}$	$86.3_{\pm 2.6}$	$85.9_{\pm 2.5}$	$86.8_{\pm 2.4}$	$87.6_{\pm 2.5}$
Multi-word	66.4% (2,380)	$58.6_{\pm 2.6}$	$57.9_{\pm 2.7}$	$60.2_{\pm 2.6}$	$59.5_{\pm 2.6}$	$60.1_{\pm 2.6}$	$62.7_{\pm 2.7}$
Document recurrence							
First occurrence	84.4% (3,023)	$66.2_{\pm 2.1}$	$65.1_{\pm 2.2}$	$67.6_{\pm 2.1}$	$67.1_{\pm 2.1}$	$67.4_{\pm 2.2}$	$69.6_{\pm 2.1}$
Recurring occurrence	15.6% (559)	$73.7_{\pm 4.5}$	$74.8_{\pm 4.2}$	$76.0_{\pm 4.1}$	$75.3_{\pm 4.3}$	$77.6_{\pm 3.8}$	$78.7_{\pm 4.1}$
2nd occurrence	11.5% (412)	$73.3_{\pm 4.2}$	$73.5_{\pm 4.7}$	$74.8_{\pm 4.4}$	$74.0_{\pm 4.3}$	$74.5_{\pm 4.0}$	$76.7_{\pm 4.2}$
3rd occurrence	2.7% (95)	$78.9_{\pm 8.0}$	$77.9_{\pm 7.6}$	$80.0_{\pm 7.4}$	$77.9_{\pm 8.3}$	$86.3_{\pm 6.1}$	$84.2_{\pm 6.7}$
4th occurrence	0.9% (34)	$67.6_{\pm 15.2}$	$76.5_{\pm 14.6}$	$79.4_{\pm 13.3}$	$79.4_{\pm 14.7}$	$85.3_{\pm 12.7}$	$85.3_{\pm 12.7}$
5th+ occurrence	0.5% (18)	$66.7_{\pm 21.9}$	$83.3_{\pm 20.6}$	$77.8_{\pm 22.2}$	$83.3_{\pm 22.2}$	$88.9_{\pm 17.6}$	$83.3_{\pm 22.2}$
Concept structure							
Simple concepts	97.9% (3,508)	$68.3_{\pm 2.2}$	$67.5_{\pm 2.1}$	$69.9_{\pm 2.1}$	$69.4_{\pm 2.1}$	$70.0_{\pm 2.1}$	$72.0_{\pm 2.0}$
Composite concepts	2.1% (74)	$21.6_{\pm 8.9}$	$25.7_{\pm 9.3}$	$24.3_{\pm 9.5}$	$17.6_{\pm 7.1}$	$21.6_{\pm 8.9}$	$24.3_{\pm 9.3}$

Table 13: Detailed subset results on MedProcNER. The first two rows report overall Recall@1 and Recall@5. All other rows report Recall@1 on specific mention subsets. The support column gives the percentage and number of examples in each subset. Each score is shown with half of the 95% confidence interval. First occurrence refers to mentions whose gold concept has not appeared earlier in the document, while recurring occurrence refers to mentions whose gold concept has already appeared earlier. Simple and composite concept rows are reported only for the Spanish SPACCC datasets.

Outcome	Example
LongBEL ✓ **Baseline** ✗	**Mention:** [ND] **Local context:** Therefore, we investigated factors associated with [ND] in patients with LAA. **Relevant global context:** In some patients with acute ischemic stroke, neurological deterioration (ND) is observed, and it is difficult to predict at the time of admission. **Previous normalized mentions:** neurological deterioration → Neurological status deteriorated ND → Neurological status deteriorated large-artery atherosclerosis → Large artery atherosclerosis **Gold:** Neurological status deteriorated (C1536136) **Baseline:** Nicotine dependence (disorder) (C0028043) ✗ **LongBEL-8B:** Neurological status deteriorated (C1536136) ✓ *Takeaway:* The abbreviation "ND" is explicitly defined earlier in the document as neurological deterioration. The sentence-level baseline only sees the ambiguous abbreviation and links it to nicotine dependence, whereas LongBEL uses the previous abbreviation definition to recover the correct concept.
LongBEL ✓ **Baseline** ✗	**Mention:** [PET] **Local context:** She was admitted overnight with suspected [PET] and was started on urgent treatment. **Relevant global context:** A 29-year-old pregnant woman presented with severe-range hypertension, headache, and epigastric pain. Laboratory testing showed proteinuria and mildly elevated liver enzymes. **Previous normalized mentions:** pregnant woman → Pregnant Woman severe-range hypertension → Hypertensive disease proteinuria → Proteinurias **Gold:** Pre-eclamptic toxemia (C0032914) **Baseline:** Petechial haemorrhage (C0031256) ✗ **LongBEL-8B:** Pre-eclamptic toxemia (C0032914) ✓ *Takeaway:* Unlike an explicit abbreviation-definition case, this example requires combining several document-level clinical cues. The local sentence only contains the ambiguous abbreviation "PET", while LongBEL uses the obstetric context and previous normalized mentions, especially "pregnant Woman", "hypertensive disease", and "proteinurias", to recover the correct concept.
LongBEL ✗ **Baseline** ✓	**Mention:** [IFN-γ] **Local context:** The aim of the study is to investigate the association of [IFN-γ] and IL-10 gene single nucleotide polymorphisms. **Global context:** (...) Interferon gamma (IFN-γ) and interleukin 10 (IL-10) polymorphisms in Chinese children (...) **Previous normalized mentions:** Interferon gamma → Interleukin 9 Gene IFN-γ → Interleukin 9 Gene interleukin 10 → Interleukin 10 IL-10 → Interleukin 10 **Gold:** IFNG gene (C1334085) **Baseline:** IFNG gene (C1334085) ✓ **LongBEL-8B:** Interleukin 9 Gene (C1334127) ✗ *Takeaway:* Although the document contains an explicit definition, LongBEL mislinks the first occurrence "Interferon gamma (IFN-γ)" and then propagates this incorrect prediction through memory to the later mention.
LongBEL ✗ **Baseline** ✗	**Mention:** [MK-8(H2)] **Local context:** The major menaquinone were [MK-8(H2)] (72%) and MK-9(H2) (28%). **Gold:** vitamin MK 8 (C0084980) **Baseline:** MAPKAPK2 protein, human (C1452441) ✗ **LongBEL-8B:** vitamin K2 (C0086605) ✗ *Takeaway:* The mention is a specialized chemical abbreviation with no additional disambiguating evidence elsewhere in the document. LongBEL predicts a related vitamin K concept, but both models fail to recover the exact KB concept.
LongBEL ✓ **Baseline** ✓	**Mention:** [DDIs] **Local context:** Elderly patients are at high risk from drug-drug interactions ([DDIs]). **Gold:** Drug Interactions (C0687133) **Baseline:** Drug Interactions (C0687133) ✓ **LongBEL-8B:** Drug Interactions (C0687133) ✓ *Takeaway:* The local sentence explicitly defines the abbreviation, so both the sentence-level baseline and LongBEL resolve the mention correctly.

Table 14: Qualitative examples comparing the sentence-level Llama-8B baseline and LongBEL-8B. The examples show when document-level context helps, when memory propagates an error, and when both systems fail or succeed. All examples are from MedMentions-ST21pv except the manually constructed PET example.

(a) ND: using an earlier abbreviation definition

low attribution — high attribution

Context

In some patients with acute ischemic stroke, neurological deterioration (ND) is observed, and it is difficult to predict at the time of admission. Especially in some patients with large-artery atherosclerosis (LAA), aggressive medical treatments and surgical interventions might be helpful to prevent ND. Therefore, we investigated factors associated with [ND] in patients with LAA.

Previous Normalizations

[acute ischemic stroke]{Disorders} Ischemic stroke
[neurological deterioration]{Disorders} Neurological status deterioration
[ND]{Disorders} Neurological status deterioration
[large-artery atherosclerosis]{Disorders} Aneurysm (morphologic abnormality)
[LAA]{Disorders} Aneurysm (morphologic abnormality)
[ND]{Disorders} Neurological status deterioration

Prediction

[ND]{Disorders} → Neurological status deterioration ✓

LongBEL-8B gives high saliency to the earlier definition "neurological deterioration (ND)" and to the corresponding memory entries. This helps the model normalize the later mention of ND correctly.

(b) PET: using clinical context and memory

Context

A 29-year-old pregnant woman presented with severe-range hypertension, headache, and epigastric pain. Laboratory testing showed proteinuria and mildly elevated liver enzymes. She was admitted overnight with suspected [PET] and was started on urgent treatment.

Previous Normalizations

[pregnant woman]{Living Beings} Pregnant Woman
[severe-range hypertension]{Disorders} Hypertensive disease
[proteinuria]{Disorders} Proteinurias

Prediction

[PET]{Disorders} → Pre-eclamptic toxemia ✓

LongBEL-8B uses the obstetric context and previous normalized mentions, especially hypertension and proteinuria, to link PET to pre-eclamptic toxemia.

(c) IFN-γ: when memory propagates an error

Context

Interferon gamma (IFN-γ) and interleukin-10 (IL-10) polymorphisms in Chinese children with hemophagocytic lymphohistiocytosis. The aim of the study is to investigate the association of [IFN-γ] and IL-10 gene single nucleotide polymorphisms with the susceptibility of hemophagocytic lymphohistiocytosis (HLH) in Chinese children without known family history of HLH. Forty children with HLH and 160 age - and gender -matched healthy controls from Xuzhou Children's Hospital were enrolled in the study. Serum IFN-γ and IL-10 levels were measured by enzyme linked-immunosorbent assay. Polymorphisms of the IFN-γ gene at position +874 and +2109, and IL-10 at position -1082 were analyzed by allele - specific PCR. Median serum concentrations of IFN -γ and IL-10 were significantly higher in children with HLH compared to healthy controls. The frequencies of IFN-γ +874 T/A and T/T genotypes, as well as T allele, were significantly higher in the HLH group compared with those in the control group. The frequencies of IL-10 -1082 G/A genotype and G allele were significantly increased in HLH patients compared with healthy controls. No significant difference was found in the distribution of IFN-γ +2109 G/A genotypes between children with HLH and controls. This study presents preliminary evidence for the association between IFN +874 T/A, T/T, IL-10 -1082 A/G genotypes, and HLH susceptibility in Chinese children with HLH.

Previous Normalizations

[Interferon gamma]{Genes & Molecular Sequences} Interleukin 9 Gene
[IFN-γ]{Genes & Molecular Sequences} Interleukin 9 Gene
[interleukin-10]{Genes & Molecular Sequences} Interleukin 10
[IL-10]{Genes & Molecular Sequences} Interleukin 10
[Chinese]{Living Beings} Chinese People
[hemophagocytic lymphohistiocytosis]{Disorders} Hemophagocytic lymphohistiocytosis
[study]{Procedures} Study

Prediction

[IFN-γ]{Genes & Molecular Sequences} → Interleukin 9 Gene ✗

LongBEL-8B gives high saliency to an incorrect previous normalization. This shows that an early linking error can influence a later abbreviation prediction through memory.

Figure 6: Saliency maps for three examples. Panels (a) and (c) are from MedMentions-ST21pv. Panel (b) is a manually constructed example. Panel (a) shows how an earlier abbreviation definition helps LongBEL-8B make the correct prediction. Panel (b) shows how clinical context and previous normalized mentions help the model resolve an ambiguous abbreviation. Panel (c) shows a failure where memory propagates an incorrect earlier prediction.

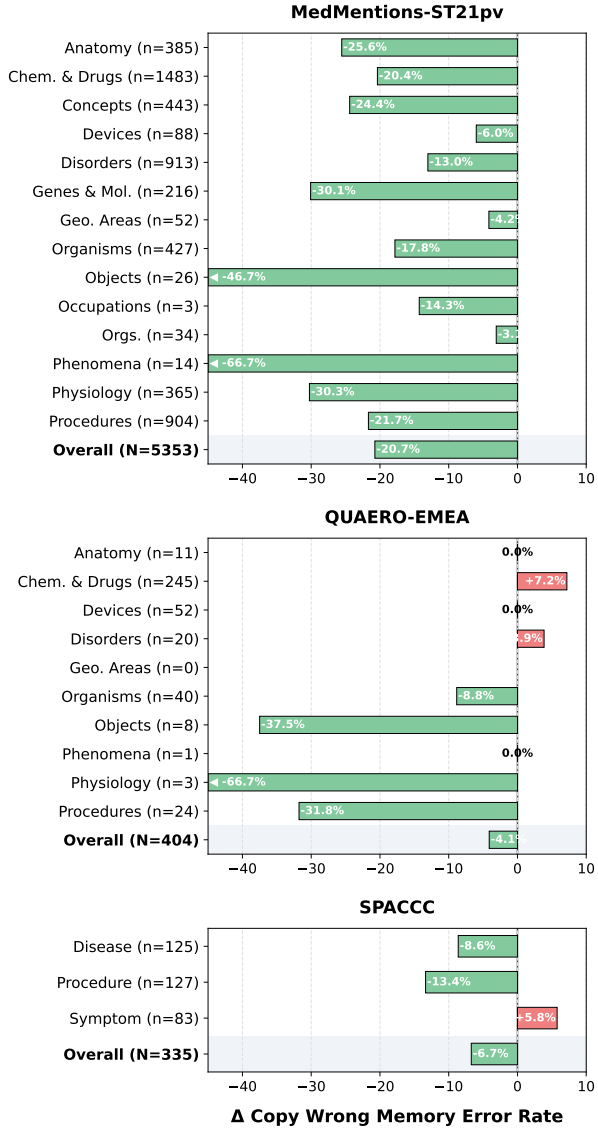

Figure 7: Change in Copy Wrong Memory Error Rate when using robust memory instead of gold memory for LongBEL-1B. Negative values indicate fewer cascading errors. Results are computed on exposed recurring concepts and reported by semantic group.

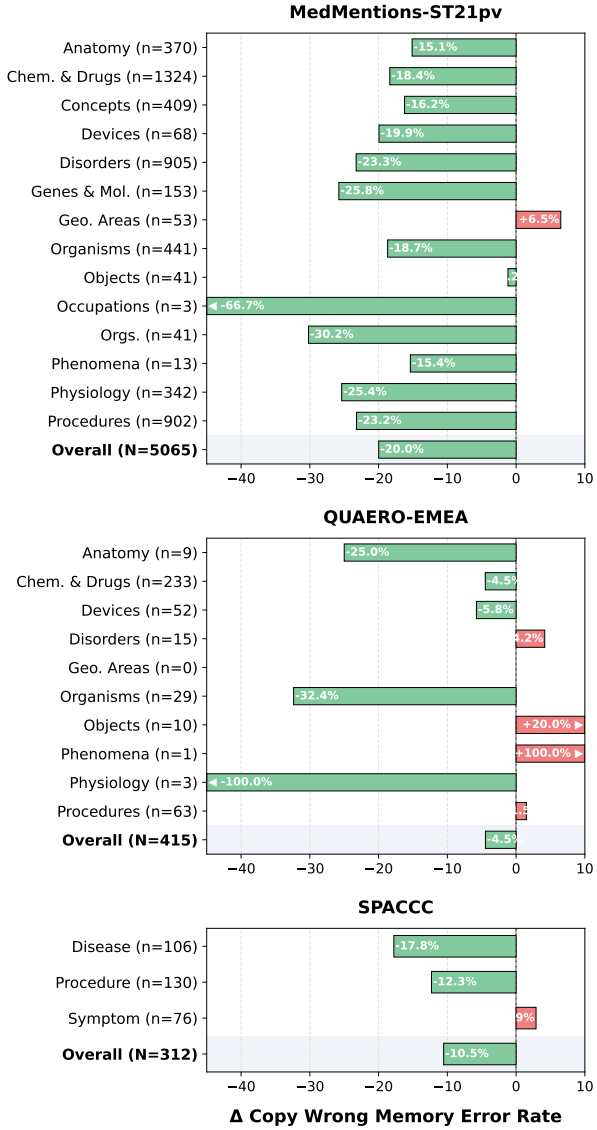

Figure 8: Change in Copy Wrong Memory Error Rate when using robust memory instead of gold memory for LongBEL-8B. Negative values indicate fewer cascading errors. Results are computed on exposed recurring concepts and reported by semantic group.