

NightSun at CLEF 2026 GutBrainIE: A First-Place System for NER, Entity Linking, and Relation Extraction

NightSun at CLEF 2026

Yeyang Liu^{1,*†}, Chunyu Li^{2,†} and Zhaohang Niu^{3,†}

¹Department of Informatics, University of Zurich, Zürich, Switzerland

²Individual Researcher

³Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya

Abstract

Gut-brain axis literature spans neuroscience, gastroenterology, microbiology, and dietetics simultaneously, demanding systems that handle heterogeneous vocabulary, overlapping entity categories, and near-synonymous relation predicates. We present NightSun, a sequential pipeline for all four GutBrainIE 2026 subtasks: BiomedBERT-CRF with relabeled bronze data and cross-model voting for NER, a SapBERT cascade with type-prior reranking for entity linking, and typed entity markers with per-class threshold sweep for relation extraction. NightSun ranks first in all four official subtasks. Per-class analysis suggests encoder-specific specialization associated with pretraining corpora, invisible to macro-F1: ModernBERT excels at microbiome, BioBERT at gene, and BioLinkBERT-large at bacteria and drug, and cross-model voting exploits this complementarity more effectively than model scaling. For entity linking, retrieval accuracy is near-ceiling for correctly detected entities whose gold concepts are present in the KB, end-to-end performance is bounded by NER errors and KB coverage. For relation extraction, NER error propagation accounts for most of the oracle-to-pipeline gap, while semantically overlapping predicates are the primary within-classifier bottleneck.

Keywords

NER, entity linking, relation extraction, biomedical NLP, gut-brain axis

1. Introduction

The gut-brain axis, the bidirectional communication network between the gastrointestinal tract and the central nervous system, has become a major focus of biomedical research, linking neurological, metabolic, and psychiatric disorders to gut microbiome composition and function [1]. The rapidly growing literature on this topic demands automated information extraction tools capable of identifying entities, grounding them to ontologies, and extracting structured relations from unstructured text.

The GutBrainIE task [2] within the BioASQ lab [3], now in its 2026 edition, defines four subtasks over biomedical abstracts from gut-brain axis research:

- **Subtask 6.1.1 - Named Entity Recognition (NER):** Identify mentions of 13 entity types in biomedical text.
- **Subtask 6.1.2 - Named Entity Recognition and Disambiguation (NERD):** Assign a canonical ontology URI to each recognized entity mention.
- **Subtask 6.2.1 - Mention-level Relation Extraction (M-RE):** Identify directed relations between entity mention pairs from 17 predicate types.
- **Subtask 6.2.2 - Concept-level Relation Extraction (C-RE):** Lift mention-level triples to canonical concept URIs.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

† These authors contributed equally.

‡ Code and trained models: <https://github.com/Kaggle-Competitions-Code/gutbrainie26>.

✉ yeyang.liu@uzh.ch (Y. Liu); li.chunyu0412@gmail.com (C. Li); niuzhaohang@gmail.com (Z. Niu)

🌐 <https://cv.yeyang.top/> (Y. Liu); <https://eeyorelee.github.io/> (C. Li)

🆔 0009-0002-3353-3236 (Y. Liu); 0009-0000-3036-0275 (C. Li); 0009-0002-9813-7955 (Z. Niu)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1

GutBrainIE 2026 dataset statistics. Rel. triples: Subtask 6.2.1 mention-level relation count.

Split	Articles	NER entities	Rel. triples
Train (gold, 2026)	639	20,530	8,556
Train (silver, 2026)	811	26,134	10,907
Train (silver, 2025)	499	15,275	10,616
Train (bronze, 2026)	2,972	89,987	29,692
Dev (2026)	80	2,521	1,139
Test (2026)	80	–	–

Test set annotations not publicly released at time of writing.

These subtasks present several compounding challenges specific to the gut-brain axis domain. First, the domain spans neuroscience, gastroenterology, microbiology, and dietetics simultaneously: a single abstract may reference microorganism taxonomy (*Lactobacillus rhamnosus* GG), dietary components (*inulin*), psychiatric disorders (*depression*), and statistical methods (*ANOVA*), each demanding different recognition signals. Second, the 13 entity categories contain semantically overlapping boundaries: microbiome and bacteria overlap when a specific bacterium is discussed in the context of gut microbiota; food and dietary supplement overlap for functional foods; DDF absorbs both diseases and clinically observed findings. Third, microbiome taxonomy is actively evolving: NCBI Taxonomy [4] and GTDB [5] often disagree on species boundaries, complicating entity disambiguation. Finally, near-synonymous predicates (*influence*, *affect*, *impact*) share surface patterns while expressing distinct semantic nuances that relation classifiers easily confuse, as we observe in our experiments.

We present NightSun, a sequential pipeline addressing all four subtasks: Subtask 6.1.1 NER entities feed Subtask 6.1.2 disambiguation and Subtask 6.2.1 mention-level relation extraction, and Subtask 6.2.1 predictions are converted to Subtask 6.2.2 concept-level triples using Subtask 6.1.2 URIs.

Our main contributions are:

- Cross-model unanimous voting over heterogeneous encoder-data combinations consistently outperforms any single model; per-class analysis suggests encoder-specific categorical advantages associated with pretraining corpora.
- A three-stage entity linking cascade (dictionary lookup → SapBERT [6] → type-prior reranking) reaches near-ceiling accuracy under the correctly detected, in-KB setting; in the full pipeline, performance is primarily bounded by NER error propagation and KB coverage.
- Typed entity markers [7] with CE-Dice loss and per-class threshold sweep effectively handle the 17-predicate class imbalance; soft probability ensembling outperforms majority voting for near-synonymous predicates.
- End-to-end error analysis shows the oracle-to-pipeline RE gap is driven by upstream NER errors, not the relation classifier, identifying NER improvement as the highest-leverage direction for Subtask 6.2.1.

Table 1 summarizes the dataset statistics for training and development splits used across all four subtasks.

2. Related Work

Biomedical NER. BERT-based models with CRF decoding [8] have become the standard for biomedical NER, with domain-specific pretraining playing a key role [9, 10, 11]. BiomedBERT [10] and BioLinkBERT [12] are among the strongest biomedical encoders for sequence labeling tasks. Sentence-boundary chunking for long abstracts is common practice when inputs exceed the 512-token limit. Multi-seed ensembling and cross-model voting improve robustness in NER. Our work adds a systematic study of training data quality, including bronze relabeling, and cross-model ensembling specific to the gut-brain axis domain.

Table 2

Training data configurations. “Samples” counts chunked title/abstract fragments (see Section 3.5).

Config	Articles	Samples
gold+silver (all)	1,949	4,146
+ original bronze	4,921	10,368
+ relabeled bronze	4,921	10,368

Biomedical entity linking. Neural entity linking methods commonly retrieve candidates from a knowledge base with dense retrieval before reranking. SapBERT [6] achieves strong performance on standard benchmarks by fine-tuning on biomedical synonym pairs, clustering synonymous concepts in embedding space. FAISS [13] enables sub-second approximate nearest-neighbor search over large KB indexes at inference time. We complement SapBERT retrieval with a type-prior reranking stage.

Relation extraction. Typed entity markers [7] outperform span-based and graph-based RE models on standard benchmarks by encoding entity type information directly into the encoder’s context. Dice loss [14] addresses the severe class imbalance in relation extraction datasets where no_relation pairs vastly outnumber positive instances; combining it with cross-entropy (CE-Dice) provides both stable gradient signal and effective minority-class focus.

3. Subtask 6.1.1 - Named Entity Recognition

3.1. Model Architecture

We frame Subtask 6.1.1 as token-level sequence labeling using a BIO tagging scheme with 27 labels, comprising O plus B- and I- prefixes for each of the 13 entity types. Our model is BertCrFForTokenClassification [15]: a pretrained encoder followed by a dropout layer, a linear emission head, and a CRF layer [8]. The default encoder is BiomedBERT-base [10]; we evaluate alternative encoders in Section 6.4.

Given an input token sequence $\mathbf{x} = (x_1, \dots, x_n)$, BiomedBERT produces contextual representations $\mathbf{h} = (\mathbf{h}_1, \dots, \mathbf{h}_n) \in \mathbb{R}^{n \times 768}$. The linear head computes emission scores $\mathbf{e}_i = \mathbf{W}\mathbf{h}_i + \mathbf{b} \in \mathbb{R}^{27}$ for each token position. The CRF models the joint probability of a label sequence $\mathbf{y}$:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp(\sum_i e_{i,y_i} + \sum_i T_{y_i,y_{i+1}})}{Z(\mathbf{x})}$$

where $T \in \mathbb{R}^{27 \times 27}$ is a learned transition matrix and $Z(\mathbf{x})$ is the partition function. Training minimizes negative log-likelihood over the CRF; inference uses the Viterbi algorithm. Each input is prefixed with a single [CLS] token. Title and abstract body are encoded as separate inputs; the title is not prepended to the abstract, preserving the full 512-token context window for the abstract text.

3.2. Training Data

GutBrainIE provides four annotation tiers in order of quality: 2025 silver, 2026 gold, 2026 silver, and 2026 bronze. Table 2 summarizes the data configurations used in our experiments.

Bronze relabeling. The bronze annotations are produced by distant supervision and contain substantial noise, particularly false positives for rare categories: gene predictions in bronze agree with gold-retrained model output only 59% of the time. We re-annotate the 2,972 bronze articles using a cross-model ensemble trained exclusively on gold+silver data: three BiomedBERT and three BioLinkBERT models vote on entity predictions, retaining only spans where $\geq 2/3$ agree. The relabeled bronze achieves 86.5% span-level agreement with the original annotations overall (Table 3). Agreement is highest for human (93.0%) and DDF (91.4%), which have distinctive surface forms; lowest for gene (58.7%) and statistical technique (69.0%), reflecting over-prediction in the original distant supervision. The relabeled bronze reduces noise while preserving the volume advantage of 2,972 additional articles.

Table 3

Per-category span-level agreement between original and relabeled bronze annotations.

Entity Type	Original	Agreement
DDF	34,560	91.4%
chemical	13,054	84.2%
microbiome	8,625	89.2%
anatomical location	6,409	79.1%
human	5,874	93.0%
bacteria	4,959	87.5%
biomedical tech.	3,851	72.0%
animal	3,090	88.5%
dietary supplement	3,017	83.4%
drug	2,358	74.8%
food	1,742	75.3%
gene	1,304	58.7%
stat. technique	1,144	69.0%
Overall	89,987	86.5%

3.3. Training

We optimize with AdamW [16] using separate learning rates: 2×10^{-5} for the BiomedBERT encoder, 2×10^{-4} for the linear head, and 2×10^{-3} for the CRF transition matrix. This three-rate scheme reflects the different convergence speeds of pretrained vs. randomly initialized components. The learning rate follows a linear decay schedule with 1% warmup. We train for up to 40 epochs with early stopping (patience 5) based on development set entity-level F1, batch size 32, weight decay 10^{-2} . This training configuration follows our prior work on the same task [17].

3.4. Class Imbalance

NER in the gut-brain axis domain faces two compounding imbalances. At the *token level*, the O label accounts for $\sim 80\%$ of all tokens even in densely annotated abstracts, with the CRF transition matrix partially mitigating this by learning to penalize implausible tag transitions. At the *entity type level*, the 13 types span a 22:1 frequency ratio in the gold+silver training data: DDF alone accounts for 33.0% of all entity instances, while food and gene are the rarest at 1.5% each.

Table 4 shows the per-type distribution alongside development set precision, recall, and F1. The correlation between training frequency and F1 is clear: high-frequency types such as DDF, chemical, and microbiome are recognized reliably, while low-frequency types such as gene, food, and statistical technique lag behind. The exception is drug: despite low training frequency, drug mentions tend to be unambiguous pharmaceutical names with distinct morphological patterns, making them easier to learn. Conversely, gene suffers from extreme ambiguity: gene symbols overlap with protein names, chemical abbreviations, and general vocabulary words, making low recall (0.365) the primary bottleneck.

The bronze data adds approximately 89,987 additional entity instances while preserving nearly the same type distribution: DDF remains dominant at 38.4%, and gene and food remain underrepresented at 1.4% and 1.9% respectively. Although rare-type absolute counts increase substantially (food: 685 $\rightarrow$ 2,427; gene: 693 $\rightarrow$ 1,997), the relative frequency imbalance persists. The largest F1 gains from adding bronze are 18.7 points on food and 5.3 points on dietary supplement over the gold+silver-only baseline, confirming that rare-category performance is primarily limited by training data volume rather than ambiguity alone. Bronze relabeling further refines these gains by reducing distant-supervision false positives: gene benefits most, as original bronze agrees with gold-trained predictions only 59% of the time, while food shows 75.3% agreement (see Table 3).

Table 4

Per-entity-type training frequency (gold+silver) and development set P/R/F1 for the primary Subtask 6.1.1 submission (run21, 3-way unanimous vote).

Entity Type	Train	Train%	P	R	F1
DDF	15,413	33.0	0.924	0.903	0.913
chemical	7,429	15.9	0.804	0.787	0.796
microbiome	4,163	8.9	0.893	0.935	0.913
human	3,786	8.1	0.947	0.922	0.934
bacteria	3,566	7.6	0.830	0.798	0.813
biomedical tech.	2,750	5.9	0.807	0.633	0.710
anatomical loc.	2,654	5.7	0.819	0.828	0.824
animal	2,112	4.5	0.933	0.921	0.927
dietary supplement	1,600	3.4	0.697	0.719	0.708
drug	1,008	2.2	0.868	0.880	0.874
stat. technique	805	1.7	0.771	0.771	0.771
gene	693	1.5	0.639	0.365	0.465
food	685	1.5	0.900	0.610	0.727

Table 5

Subtask 6.1.1 seed stability (Micro-F1) across seeds 7, 11, 17. Dev Ens.: 3-seed within-model ensemble (avg. emission + CRF). Test Ens.: official leaderboard Micro-F1 for the corresponding submitted ensemble run.

Base Model	Train Data	Dev Min	Dev Max	Dev Ens.	Test Ens.
BiomedBERT	gold+silver (gs)	0.840	0.844	0.847	0.835
BiomedBERT	gs + original bronze	0.836	0.842	0.844	0.836
BiomedBERT	gs + relabeled bronze	0.844	0.848	0.852	0.838
BioLinkBERT	gold+silver	0.833	0.839	0.842	0.824

3.5. Chunked Inference and Ensembling

Long abstracts are split into non-overlapping chunks of at most 512 tokens; chunk boundaries are placed at the last punctuation mark before the token limit to avoid mid-sentence splits. Each chunk is processed independently and predictions are merged back to the original character offsets. This ensures entities near the end of long abstracts are recovered, which standard truncation would miss.

We train three seeds (7, 11, 17) per configuration and construct *within-model ensembles* by averaging emission score matrices and CRF transition matrices across the three seeds before Viterbi decoding. This soft average operates at the score level and consistently outperforms majority voting on predictions.

Table 5 reports seed stability across configurations (Micro-F1). The relabeled-bronze BiomedBERT achieves the best single-seed Micro-F1, confirming that cleaner training data reduces optimization instability. The original bronze shows the highest variance due to inconsistent annotation noise. The 3-seed ensemble consistently exceeds the best single seed on Micro-F1 across all configurations.

The best Subtask 6.1.1 submission (run21) applies *cross-model unanimous voting* (3/3 agreement) across three within-model ensembles: BiomedBERT trained on gold+silver, on original bronze, and on relabeled bronze. All three share the BiomedBERT encoder but differ in training data, spanning the quality spectrum from curated annotations to relabeled distant supervision. For downstream Subtask 6.1.2 and Subtasks 6.2.1/6.2.2, we use run06 (3-way majority vote, $\geq 2/3$ agreement) as the NER source, as its higher recall reduces entity miss rate in the pipeline at acceptable precision cost.

4. Subtask 6.1.2 - Named Entity Recognition and Disambiguation

4.1. Knowledge Base Construction

We build a unified knowledge base (KB) from ontologies spanning all 13 entity types. The KB contains 52,263 concept aliases derived from MeSH [18], NCBI Gene [19], GTDB [5] and NCBI Taxonomy [4], FoodOn [20] and USDA FNDDS [21], and the URI set from the task’s training annotations.

For each concept, we index all name variants as separate entries. A direct lookup table of 18,512 (surface form, entity type) $\rightarrow$ URI mappings is extracted from the training and development annotations for high-precision exact matching.

4.2. Three-Stage Linking Pipeline

Stage 1: Dictionary lookup. We first query the 18,512-entry lookup table by (normalized surface form, entity type). Normalization includes lowercasing, punctuation normalization, and common abbreviation expansion. Dictionary lookup handles common, unambiguous mentions with near-perfect precision and processes all test entities in under a second.

Stage 2: SapBERT dense retrieval. For entities not resolved by dictionary lookup, we encode the mention string with SapBERT [6] (cambridgelt1/SapBERT-from-PubMedBERT-fulltext) and retrieve the top- k ($k = 20$) candidate URIs from a FAISS [13] index over the 52,263 KB alias embeddings (768-dimensional). SapBERT is pretrained on the UMLS synonym graph using a metric-learning objective, placing synonymous biomedical concepts close together in embedding space. Conditioned on correct entity detection and KB coverage, SapBERT achieves 95.6% top-1 accuracy.

Stage 3: Type-prior reranking. The top- k SapBERT candidates are reranked by a linear combination of similarity score and type-conditioned frequency prior:

$$\text{score}(c) = \text{sim}(e, c) + \alpha \cdot \log p(c \mid t)$$

where t is the entity type and $p(c \mid t)$ is the empirical probability of concept c for type t in the training annotations. We set $\alpha = 0.05$ based on a development set sweep; given SapBERT’s high baseline accuracy (95.6% top-1), the gain from reranking is marginal, but it corrects type-ambiguous surface forms (e.g., a chemical that also appears as a drug) without over-relying on frequency priors for rare entities.

5. Subtask 6.2 - Relation Extraction

5.1. Model Architecture

For Subtask 6.2.1, we fine-tune BiomedBERT-base for 18-class relation classification (no_relation + 17 legal predicates) using typed entity markers [7]. For a candidate pair with subject entity of type t_s and object entity of type t_o , markers are inserted into the article text:

- Subject: @ * t_s * span @
- Object: # t_o span #

The classifier reads the hidden states at the @ and # marker positions, concatenates them, and applies a linear head. This design keeps type information in the encoder’s receptive field and avoids separate entity representations.

5.2. Training Data and Loss

Positive pairs come from the 2026 gold and 2025 silver annotations; the development set contains 1,139 GT mention-level relations spanning 17 predicates. The predicate distribution is highly skewed: *located in* (196), *influence* (160), and *target* (144) together account for 44% of all relations, while *compared to*

(4) and *produced by* (6) together contribute under 1%. Negative (no_relation) pairs are sampled from all type-valid entity pairs in the same article at ratios of 5:1 ($\lambda_{CE} = 0.5$ variants) and 7:1 ($\lambda_{CE} = 0.8$ variants); type validity is determined by the set of (subject type, object type) pairs observed in training. This constraint is important for efficiency: the 13 entity types yield 169 ordered (subject, object) type pairs, but only 54 distinct combinations appear in the training annotations. Restricting to type-valid pairs substantially reduces the candidate count, yielding 50,091 candidate pairs on the development set (gold NER), while retaining all gold relation instances (no gold relation pair violates the type constraint by definition).

The severe class imbalance, where no_relation accounts for $\approx 87\%$ of training pairs in the 7:1 configuration and the rarest predicate *compared to* has only 4 development instances, motivates a combined loss:

$$\mathcal{L} = \lambda_{CE} \cdot \mathcal{L}_{CE} + (1 - \lambda_{CE}) \cdot \mathcal{L}_{Dice}$$

The cross-entropy component $\mathcal{L}_{CE}$ provides stable gradient signal; the Dice loss [14] component down-weights easy negatives and focuses learning on under-represented predicates. We train two CE/Dice ratio variants: $\lambda_{CE} \in \{0.5, 0.8\}$, combined with three seeds $\{7, 11, 17\}$, yielding six trained RE models in total.

For probability and logit averaging, we use the five models included in the best development-set ensemble, excluding the remaining model because it reduced development micro-F1 during ensemble selection. The majority-vote baseline uses all six models with a $\geq 4/6$ agreement threshold.

Training uses AdamW with learning rate 1×10^{-4} (classifier head), 2×10^{-5} (BERT base), batch size 32, max 30 epochs, early stopping (patience 5) on development micro-F1.

5.3. Windowed Tokenization

Standard 512-token truncation silently removes entity markers for long articles, causing the classifier to see incomplete context or, worse, a truncated marker. We implement windowed tokenization: for sequences exceeding 512 tokens, a window of $512 - 2 = 510$ content tokens is extracted, centered on the midpoint between the two marker tokens. The [CLS] and [SEP] boundaries are preserved. If the span between the two markers alone exceeds 510 tokens, the pair is skipped, as no context window can simultaneously contain both. On the test set, standard truncation would drop 1,926 of 61,593 candidate pairs (3.1%); windowed tokenization recovers all of them.

5.4. Subtask 6.2.2 - Concept-level Relation Extraction

Subtask 6.2.2 predictions are derived from Subtask 6.2.1 by replacing each entity mention span with its canonical URI from the Subtask 6.1.2 NERD predictions. For each Subtask 6.2.1 triple (e_s, p, e_o) , we look up the (text, label) pair for both entities in the Subtask 6.1.2 output; triples where either entity lacks a URI are discarded. This yields 85% coverage of Subtask 6.2.1 mention-level triples as concept-level triples. An additional deduplication step is applied: multiple Subtask 6.2.1 mention triples that map to the same (uri_s, p, uri_o) concept triple are collapsed into one, which slightly reduces the total count but avoids redundant concept-level assertions. On the test set, the best Subtask 6.2.2 submission, run18 derived from Subtask 6.2.1 run18, contains 1,369 concept-level triples from 1,608 mention-level source relations, with 239 triples removed either from missing URIs or deduplication.

6. Experimental Results

6.1. Subtask 6.1.1 - Named Entity Recognition

Tables 6 and 7 present Subtask 6.1.1 results on the 2026 development set. The best submission, run21, uses a 3-way unanimous vote across three within-model BiomedBERT ensembles trained on different data configurations: gold+silver, gold+silver+original bronze, and gold+silver+relabeled bronze. It achieves **macro-F1 0.798**, **micro-F1 0.854** on the development set and **macro-F1 0.769**, **micro-F1**

Table 6

Subtask 6.1.1 NER Micro-F1. Dev: 2026 development set. Test: official leaderboard. **Bold**: primary submission. Data: gs = gold+silver; gs+orig = gs + original bronze; gs+relab = gs + relabeled bronze; all = all three data variants.

Ensemble	Data	Encoder	Dev Mi	Test Mi
<i>Cross-model ensembles</i>				
3-way unanimous	all	BiomedBERT	0.854	0.842
3-way majority	all	BiomedBERT	0.853	0.838
5-way vote	all	BiomedBERT+BioLink	0.854	0.837
<i>Within-model 3-seed ensembles</i>				
3-seed	gs	BiomedBERT	0.847	0.835
3-seed	gs+orig	BiomedBERT	0.844	0.836
3-seed	gs+relab	BiomedBERT	0.852	0.838
3-seed	gs	BioLinkBERT	0.842	0.824
3-seed	gs+relab	BioLinkBERT	0.843	0.828

Table 7

Subtask 6.1.1 per-entity results for the primary submission (run21, 3-way unanimous vote, dev set).

Entity Type	P	R	F1	GT
human	0.947	0.922	0.934	192
animal	0.933	0.921	0.927	152
microbiome	0.893	0.935	0.913	231
DDF	0.924	0.903	0.913	793
drug	0.868	0.880	0.874	75
anat. location	0.819	0.828	0.824	169
bacteria	0.830	0.798	0.813	183
chemical	0.804	0.787	0.796	366
stat. technique	0.771	0.771	0.771	35
food	0.900	0.610	0.727	59
biomed. tech.	0.807	0.633	0.710	139
dietary suppl.	0.697	0.719	0.708	64
gene	0.639	0.365	0.465	63

0.842 on the official test set, ranking first in Subtask 6.1.1. Unanimous voting raises precision by filtering out predictions rejected by any single model, accepting a modest recall trade-off versus majority vote ($\geq 2/3$).

Per entity type in Table 7, high-frequency types are recognized reliably: human, animal, and DDF/microbiome all exceed F1 0.9. Challenging types include gene due to ambiguity with protein and chemical names, food limited by compositional phrases such as “high-fat diet”, and biomedical technique hampered by diverse experimental method names.

6.2. Per-Class Model Comparison

Table 8 presents per-entity-type F1 for all nine single-model variants and the final 3-way vote ensemble. No single model dominates every category, which motivates cross-model ensembling and suggests encoder-specific strengths associated with pretraining corpora.

Several findings stand out. **Pretraining corpus is associated with category-specific strengths.** ModernBERT-bioclinical [22] achieves the highest microbiome F1 despite its weakest overall macro-F1; its clinical note pretraining, rich in gut microbiota terminology, is consistent with a specific advantage for microbiome mentions. BioBERT v1.1 [9] leads on gene, outperforming BiomedBERT by over 5 points; its PubMed+PMC pretraining covers full-text articles including supplementary gene tables, which may provide broader coverage of gene symbol variants. SciBERT [11] matches BiomedBERT on statistical technique, which is consistent with scientific literature pretraining that includes methods

Table 8

Per-entity-type F1 on the 2026 development set across all models and the final 3-way vote ensemble.

Bold: highest F1 per category. All single models use seed=11, chunked inference. BiomedBERT columns: gold+silver only; +brz: gold+silver+original bronze; +rlb: gold+silver+reabeled bronze.

Label	BiomedBERT	+brz	+rlb	BioLink	BiomedBERT-L	BioLink-L	BioBERT	SciBERT	ModernBERT	3-way vote
DDF	0.907	0.886	0.897	0.898	0.900	0.898	0.885	0.877	0.886	0.913
chemical	0.787	0.799	0.786	0.791	0.787	0.788	0.767	0.728	0.693	0.796
microbiome	0.906	0.898	0.895	0.884	0.883	0.906	0.873	0.861	0.931	0.913
bacteria	0.815	0.824	0.806	0.821	0.786	0.839	0.760	0.803	0.739	0.813
anat. location	0.820	0.800	0.821	0.814	0.807	0.821	0.791	0.794	0.776	0.824
human	0.932	0.908	0.929	0.926	0.943	0.921	0.922	0.913	0.905	0.934
animal	0.931	0.929	0.925	0.931	0.920	0.861	0.923	0.906	0.886	0.927
biomed. tech.	0.653	0.696	0.700	0.667	0.674	0.647	0.620	0.618	0.677	0.710
drug	0.873	0.861	0.859	0.873	0.811	0.890	0.854	0.854	0.816	0.874
dietary suppl.	0.691	0.744	0.734	0.681	0.696	0.676	0.691	0.643	0.632	0.708
food	0.646	0.833	0.759	0.639	0.653	0.500	0.667	0.587	0.505	0.727
stat. technique	0.779	0.741	0.744	0.725	0.738	0.747	0.701	0.779	0.703	0.771
gene	0.491	0.517	0.491	0.480	0.513	0.442	0.545	0.438	0.414	0.465
Macro-F1	0.787	0.803	0.796	0.779	0.778	0.764	0.769	0.754	0.736	0.798
Micro-F1	0.843	0.842	0.844	0.839	0.836	0.834	0.821	0.811	0.804	0.854

sections with diverse statistical terminology.

Large models provide targeted gains. BiomedBERT-large achieves the best human F1, slightly exceeding the 3-way vote ensemble, while BioLinkBERT-large leads on bacteria and drug. These large-model gains are not captured by our 3-way vote ensemble, which uses only base-scale models. For bacteria, citation-link pretraining [12] through microbiology and genomics literature likely provides richer taxonomic context; for drug, the citation graph connects pharmacology papers with consistent drug name usage. However, these advantages are category-specific: large models underperform their base counterparts on macro-F1 overall, making them impractical as standalone replacements.

Bronze data helps food and dietary supplement but the unanimous threshold suppresses the gains. BiomedBERT+bronze achieves the highest food F1, well above BiomedBERT, with similar gains on dietary supplement. However, the 3-way unanimous vote drops both food and dietary supplement below the bronze-trained single model, because the gold+silver BiomedBERT model has substantially lower food F1 and thus eliminates mentions that both bronze-trained variants detect. A weighted ensemble or a lower vote threshold for rare bronze-boosted categories would recover these gains.

The 3-way vote improves over all single models on most categories. The vote ensemble surpasses every individual model on DDF, anatomical location, and biomedical technique by combining complementary errors. Gene falls below all vote-constituent models because the unanimous threshold amplifies recall losses on an already recall-limited category.

6.3. Subtask 6.1.1 Ablation Study

Table 9 quantifies the contribution of each design choice. All configurations use BiomedBERT-base and chunked inference; ablation rows use seed 11, and the Best row shows the best single-seed result (seed 7).

Original bronze adds meaningful macro-F1; relabeling it adds further gains with notably lower seed variance.

Table 9

Subtask 6.1.1 ablation study on the 2026 development set (BiomedBERT-base, seed 7 for Best; seed 11 for ablation rows).

Configuration	Ma-F1	Mi-F1
gs + relabeled bronze (Best)	0.806	0.848
gs + original bronze	0.803	0.842
gs only (no bronze)	0.787	0.843
gs + relabeled bronze, BIOES	0.776	0.840

Table 10

Subtask 6.1.1 encoder comparison (development set, gold+silver training, chunked inference).

Encoder	Params	Ma-F1	Mi-F1
<i>Base models (110–150M)</i>			
BiomedBERT-base	110M	0.787	0.843
BioLinkBERT-base	110M	0.779	0.839
BioBERT v1.1	110M	0.769	0.821
SciBERT	110M	0.754	0.811
ModernBERT-bioclinical	150M	0.736	0.804
<i>Large models (340M, batch 16 + grad. accum. 2)</i>			
BiomedBERT-large	340M	0.781	0.835
BioLinkBERT-large	340M	0.761	0.835
<i>Non-BERT</i>			
DeBERTa-v2 + CRF	184M	0.658	0.751

6.4. Subtask 6.1.1 Encoder Comparison

Table 10 compares seven biomedical BERT variants as drop-in encoder replacements under identical training and hyperparameters (gold+silver data, no bronze).

BiomedBERT-base achieves the best overall performance. BioLinkBERT-base [12] performs marginally lower, suggesting that citation-graph pretraining does not yield an overall advantage under this configuration. Scaling to 340M parameters degrades performance, as reduced batch sizes harm optimization and the larger model overfits on $\sim 4K$ training samples. DeBERTa-v2 [23] performs substantially worse, suggesting disentangled attention is less compatible with the linear-CRF decoder and BIO tagging scheme.

6.5. Subtask 6.1.2 - Named Entity Recognition and Disambiguation

The three-stage cascade achieves **micro-F1 0.787**, **macro-F1 0.731** on the 2026 development set in end-to-end pipeline mode using Subtask 6.1.1 run06 NER output. With gold NER entities, oracle linking accuracy is 91.6%: 8.4% of gold entities have no matching KB entry, including novel gene synonyms, trade names, and emerging terminologies, which is the hard ceiling for the disambiguation subtask. Conditioned on correct entity detection and KB coverage, SapBERT top-1 accuracy is 95.6%, with remaining errors dominated by annotator inconsistency where multiple valid URIs could apply.

Type-prior reranking resolves type-ambiguous surface forms without degrading rare-entity performance; higher α values monotonically decrease oracle accuracy as the prior overrides SapBERT similarity for low-frequency concepts. A cross-encoder reranker provided negligible improvement at substantially higher inference cost; substring matching also consistently underperformed SapBERT; both were abandoned.

The out-of-KB gap is concentrated in two categories. Gene entities contribute disproportionately: gene names are frequently non-standard, including unpublished gene IDs, organism-specific aliases, and

Table 11

Subtask 6.2.1 mention-level RE on the 2026 development set (pipeline mode).

Configuration	P	R	Mi-F1	Ma-F1
Oracle (gold entities)	0.571	0.675	0.619	0.610
5-model prob, $\lambda = 0$	0.460	0.515	0.486	0.465
5-model logit, $\lambda = 0$	0.428	0.544	0.479	0.453
3-model ($\lambda_{CE}=0.5$ only)	0.430	0.526	0.473	0.443
Single model	0.392	0.509	0.443	0.381
Vote ($\geq 4/6$)	0.415	0.495	0.452	0.415

Table 12

Subtask 6.2.1 per-predicate results for the best ensemble (pipeline, dev set). GT = ground truth count.

Predicate	P	R	F1	GT
target	0.585	0.715	0.644	144
is a	0.552	0.787	0.649	47
is linked to	0.650	0.609	0.629	110
change expr.	1.000	0.571	0.727	7
strike	0.583	0.609	0.596	23
change effect	0.543	0.613	0.576	31
produced by	0.417	0.833	0.556	6
affect	0.455	0.480	0.467	125
located in	0.393	0.561	0.462	196
influence	0.453	0.394	0.421	160
administered	0.337	0.472	0.393	72
used by	0.439	0.342	0.385	73
impact	0.309	0.439	0.362	57
compared to	1.000	0.250	0.400	4
interact	0.277	0.342	0.306	38
part of	0.714	0.167	0.270	30
change abund.	0.200	0.125	0.154	16

protein-family umbrella terms. Microbiome entities are the second major source: newly described taxon names or GTDB-style names diverging from NCBI Taxonomy have no KB match. For these entities, the system assigns an empty URI per the task guidelines, contributing a fixed linking precision of 0.

The overall NERD micro-F1 gap relative to oracle largely reflects NER error propagation: entities missed or incorrectly typed by Subtask 6.1.1 cannot be correctly disambiguated regardless of linker quality. Controlling for both correct entity detection and KB coverage, disambiguation accuracy is near-ceiling; therefore, the official end-to-end Subtask 6.1.2 score should not be interpreted as pure retrieval accuracy.

6.6. Subtasks 6.2 - Relation Extraction

Table 11 shows Subtask 6.2.1 results on the development set in pipeline mode using Subtask 6.1.1 run06 NER output. The best 5-model ensemble with probability averaging and $\lambda = 0$ achieves **micro-F1 0.486, macro-F1 0.465**. The oracle upper bound using gold NER entities reaches micro-F1 0.619, with the gap quantifying NER error propagation.

Per-predicate analysis (Table 12) reveals systematic patterns. Semantically specific predicates such as *target*, *is linked to*, and *is a* are recognized well. Semantically overlapping predicates are harder: *influence*, *affect*, and *impact* share similar surface patterns and are frequently confused with each other. *Part of* and *change abundance* are the weakest; *part of* shows high precision but very low recall, suggesting the model defaults to no_relation for most instances. For very rare predicates such as *change expression*, *produced by*, and *compared to*, the per-class threshold sweep successfully calibrates decision boundaries

Table 13

Official GutBrainIE 2026 test set results (best run per subtask). NightSun ranked first in all four subtasks.

Task	Best run	Ma-P	Ma-F1	Mi-P	Mi-F1
Subtask 6.1.1 NER	run21	0.823	0.769	0.876	0.842
Subtask 6.1.2 NERD	run16	0.538	0.521	0.629	0.617
Subtask 6.2.1 M-RE	run18	0.377	0.336	0.446	0.453
Subtask 6.2.2 C-RE	run18	0.200	0.189	0.248	0.263

despite minimal training signal.

Soft ensembling outperforms majority voting: probability averaging preserves the fine-grained signal needed to distinguish semantically similar predicates.

6.7. Official Test Set Results

Table 13 reports official test set scores for all four subtasks.

7. Discussion

Four cross-cutting findings emerge from our experiments.

Bronze relabeling generalizes cleanly. Re-annotating noisy distant-supervision data with a cross-model ensemble improves both NER quality and training stability, with gains concentrated in categories where distant supervision is least reliable: gene, food, and dietary supplement.

Entity linking is bounded by NER propagation and KB coverage, not only retrieval. SapBERT retrieval accuracy is near-ceiling only after conditioning on two prerequisites: the entity mention is correctly detected and typed by Subtask 6.1.1, and the correct concept exists in the KB. The official end-to-end Subtask 6.1.2 score is lower because it also includes missed entities, mistyped entities, and out-of-KB terminology. The large dev-to-test drop is therefore more consistent with increased out-of-KB and pipeline errors than with dense retrieval degradation alone. Extending the KB and improving upstream NER are the highest-impact improvements for Subtask 6.1.2.

NER quality bounds RE performance. The oracle-to-pipeline gap is consistent with the theoretical estimate derived from NER pair recall: the RE model itself is not the limiting factor. Improving Subtask 6.2.1 requires reducing NER errors upstream, not enhancing the RE model alone.

Semantically overlapping predicates are the main RE bottleneck. A small cluster of high-frequency predicates dominates GT relation counts, and *influence*, *affect*, and *impact* cause substantial mutual confusion despite covering a large share of relations. Coarsening the predicate hierarchy or contrastive training on this cluster would directly address the bottleneck.

8. Conclusion

NightSun ranked first in all four GutBrainIE 2026 subtasks.

For Subtask 6.1.1 NER, BiomedBERT-CRF with relabeled bronze data and 3-way cross-model voting achieves strong development performance; bronze relabeling is the single highest-impact design choice, substantially reducing seed variance.

For Subtask 6.1.2 NERD, a three-stage linking cascade reaches near-ceiling retrieval accuracy under the correctly detected, in-KB setting; the end-to-end gap is driven by NER error propagation and out-of-KB terminology rather than retrieval quality alone.

For Subtasks 6.2.1/6.2.2 RE, typed entity markers, CE-Dice loss, and per-class threshold sweep achieve the top-ranked Subtask 6.2.1 test performance. NER error propagation and influence/affect/impact predicate confusion are the dominant bottlenecks.

Per-class analysis suggests category-level encoder specialization associated with pretraining corpora, invisible to macro-F1: ModernBERT leads on microbiome, BioBERT on gene, and BioLinkBERT-large on bacteria and drug. Exploiting this complementarity via per-category model weighting is a natural extension of the current voting ensemble.

Future directions include jointly modeling NER and RE to reduce error propagation, expanding the disambiguation KB with community-curated synonym sets for evolving microbiome taxonomy, and exploring instruction-tuned LLMs for zero-shot generalization to rare predicates.

Declaration on Generative AI

During the preparation of this work, the authors used Claude as a writing assistance tool to paraphrase and reword text, improve writing style, and check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] J. F. Cryan, K. J. O’Riordan, C. S. Cowan, K. V. Sandhu, T. F. Bastiaanssen, M. Boehme, M. G. Codagnone, S. Cusotto, C. Fulling, A. V. Golubeva, et al., The microbiota-gut-brain axis, *Physiological reviews* 99 (2019) 1877–2013.
- [2] M. Martinelli, V. Bonato, G. M. Di Nunzio, G. Faggioli, N. Ferro, O. Irrera, B. Kantz, S. Marchesin, L. Menotti, S. Merlo, R. Michelotto, G. Tussardi, F. Vezzani, P. Waldert, G. Silvello, Overview of GutBrainIE@CLEF 2026: Gut-Brain Interplay Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), *CLEF 2025 Working Notes*, 2026.
- [3] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodríguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of Lecture Notes in Computer Science*, Springer, 2026.
- [4] C. Schoch, S. Ciufo, C. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. McVeigh, K. O’Neill, B. Robbertse, S. Sharma, V. Soussov, J. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, Ncbi taxonomy: A comprehensive update on curation, resources and tools, *Database* 2020 (2020). doi:10.1093/database/baaa062.
- [5] D. Parks, P.-A. Chaumeil, A. Mussig, C. Rinke, M. Chuvochina, H. Philip, Gtdb release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes, *Nucleic Acids Research* 54 (2025) D743–D754. doi:10.1093/nar/gkaf1040.
- [6] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics*, 2021, pp. 4228–4238. URL: <https://doi.org/10.18653/v1/2021.naacl-main.334>. doi:10.18653/V1/2021.NAACL-MAIN.334.
- [7] Z. Zhong, D. Chen, A frustratingly easy approach for joint entity and relation extraction, *CoRR abs/2010.12812* (2020). URL: <https://arxiv.org/abs/2010.12812>. arXiv:2010.12812.
- [8] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: C. E. Brodley, A. P. Danyluk (Eds.), *Proceedings of*

- the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, Morgan Kaufmann, 2001, pp. 282–289.
- [9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinform.* 36 (2020) 1234–1240. URL: <https://doi.org/10.1093/bioinformatics/btz682>. doi:10.1093/BIOINFORMATICS/BTZ682.
 - [10] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, *CoRR abs/2007.15779* (2020). URL: <https://arxiv.org/abs/2007.15779>. arXiv:2007.15779.
 - [11] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics*, 2019, pp. 3613–3618. URL: <https://doi.org/10.18653/v1/D19-1371>. doi:10.18653/v1/D19-1371.
 - [12] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics*, 2022, pp. 8003–8016. URL: <https://doi.org/10.18653/v1/2022.acl-long.551>. doi:10.18653/v1/2022.ACL-LONG.551.
 - [13] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, *CoRR abs/1702.08734* (2017). URL: <http://arxiv.org/abs/1702.08734>. arXiv:1702.08734.
 - [14] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced NLP tasks, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics*, 2020, pp. 465–476. URL: <https://doi.org/10.18653/v1/2020.acl-main.45>. doi:10.18653/v1/2020.ACL-MAIN.45.
 - [15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics*, 2019, pp. 4171–4186. URL: <https://doi.org/10.18653/v1/n19-1423>. doi:10.18653/v1/N19-1423.
 - [16] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net*, 2019. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
 - [17] J. Han, Y. Liu, Gutuzh at CLEF2025 bioasq task 6: a method of SOTA performance with the best results at gutbrainie NER subtask 1, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org*, 2025, pp. 302–316. URL: https://ceur-ws.org/Vol-4038/paper_20.pdf.
 - [18] National Library of Medicine, Medical Subject Headings, <https://www.nlm.nih.gov/mesh/meshhome.html>, 2026. Last reviewed: January 21, 2026. Accessed: 2026-07-06.
 - [19] NCBI Gene, Gene [Internet], <https://www.ncbi.nlm.nih.gov/gene/>, 2004. 2004–. Accessed: 2026-07-06.
 - [20] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, W. W. L. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, *npj Science of Food* 2 (2018). URL: <https://www.nature.com/articles/s41538-018-0032-6>. doi:10.1038/s41538-018-0032-6.
 - [21] U.S. Department of Agriculture, Agricultural Research Service, USDA Food and Nutrient Database for Dietary Studies 2021–2023, Food Surveys Research Group Home Page. <https://www.ars.usda.gov/nea/bhnrc/fsrg>, 2024. Accessed: 2026-07-06.
 - [22] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas,

- F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, Association for Computational Linguistics, 2025, pp. 2526–2547. URL: <https://aclanthology.org/2025.acl-long.127/>.
- [23] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced BERT with disentangled attention, CoRR abs/2006.03654 (2020). URL: <https://arxiv.org/abs/2006.03654>. arXiv:2006.03654.