HSE NLP Team at BioASQ BioNNE-R: When Dev Lies: Domain Pretraining, Calibration, and the Limits of LLM Augmentation for Biomedical Nested Relation Extraction

Notebook for the BioASQ BioNNE-R Lab at CLEF 2026

Airat Valiev[1,*]

[1]HSE University / Moscow, Russia

Abstract

We describe our system for the BioNNE-R 2026 shared task on nested biomedical relation extraction, achieving macro F1 of **0.4733** (English, Subtask 1), **0.5057** (Russian, Subtask 2), and **0.4527** (Bilingual, Subtask 3) on the blind test set. The system fine-tunes domain-appropriate transformer encoders (BiomedBERT for English, mDeBERTa-v3-base for the Russian and bilingual tracks) and adds three components: (1) typed entity markers with a nesting-aware, right-to-left insertion order and a binary nesting flag, which correctly encode the $\approx 15\%$ of entity pairs whose spans overlap; (2) a two-stage generate-then-verify LLM augmentation pipeline for rare relation classes; and (3) a blind-development confidence-threshold sweep that corrects the $\approx 30\times$ mismatch between the training negative ratio (3:1) and the test-time ratio induced by exhaustive entity-pair enumeration ($\approx$90:1). We also report a sharp development-to-test inversion. The model with the higher development score (augmented mDeBERTa, 0.7902) loses on the test set to a domain-specific model that scores lower on development (BiomedBERT, $0.7369 \rightarrow 0.4733$ vs. 0.3334), which we trace to calibration degradation from LLM-generated training data. We document five negative findings (augmentation harm for Russian, MC-dropout ensembles, an existence filter, per-class thresholds, and cross-lingual transfer) and draw practical guidelines for future biomedical relation-extraction systems.

Keywords

Relation extraction, Biomedical NLP, Nested named entities, Data augmentation, Confidence calibration, BioNNE-R

1. Introduction

Relation extraction (RE) identifies semantic relationships between named entities in text. It is a core component of biomedical information-extraction pipelines, underpinning applications from knowledge-graph construction to pharmacovigilance and clinical decision support. The predominant formulation treats RE as sentence-level classification over pairs of pre-annotated entity mentions. The BioNNE-R 2026 shared task at BioASQ departs from the usual flat-entity setting in a structurally important way: entities are *nested*, meaning that one entity span may contain another within its boundaries. A disorder mention such as *atopic dermatitis* (DISO) structurally contains *dermatitis* (DISO). Relations are defined not only between top-level entities but across nesting levels, so that a relation may hold between an inner entity and the entity that lexically contains it. Although nested entity recognition is itself well studied [1], relation extraction *over* nested spans has received far less attention. Common entity-encoding strategies such as masking spans, inserting boundary tokens, or pooling span representations break silently when applied to overlapping spans without explicit nesting-aware handling, making this a non-trivial extension of an otherwise well-studied problem.

The task operates over NEREL-BIO [2], a corpus of biomedical PubMed abstracts annotated in both English and Russian with eight entity types and fourteen directed relation classes. Three tracks are defined: monolingual English (Subtask 1), monolingual Russian (Subtask 2), and a bilingual track

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

*Corresponding author.

aa.valiev@hse.ru (A. Valiev)

0009-0007-1374-2002 (A. Valiev)

requiring a single model trained on combined data (Subtask 3). The multilingual dimension is practically significant: Russian biomedical NLP remains underresourced relative to English, with far fewer domain-specific pre-trained models and benchmarks. The English training partition is also small (56 documents, 3,193 annotated relations), creating a low-resource scenario inside an otherwise high-resource language, a combination that tests the limits of both domain adaptation and data augmentation.

We submitted to all three subtasks, reaching macro-averaged F1 of **0.4733** (English), **0.5057** (Russian), and **0.4527** (Bilingual) on the blind test set. [1] The paper makes four contributions:

1. **Typed entity markers with nesting-aware interleaving.** We replace the generic boundary tokens of the organiser baseline with typed markers of the form `[HEAD:DISO]` / `[/HEAD:DISO]`, encoding entity type directly into the input. For nested span pairs we introduce a right-to-left marker insertion order that preserves all character offsets, together with a binary nesting flag appended to the entity representation before classification.
2. **Two-stage LLM augmentation for rare classes.** For relation types with low training coverage and poor development performance, notably ALTERNATIVE_NAME (95 training instances, dev F1 0.2482) and APPLIED_TO (43 instances, dev F1 0.4835), we use a generate-then-verify pipeline, producing 917 English and 874 Russian augmented examples.
3. **Blind-development confidence calibration.** Exhaustive within-document pair enumeration produces $\approx$818K candidate pairs for the English test set, an effective negative ratio of $\approx$91:1, about $30\times$ the 3:1 ratio used in training. A threshold sweep over the blind development set corrects this mismatch and is the single highest-impact inference-time intervention.
4. **A systematic catalogue of negative findings.** We document five directions that did not improve test performance: MC-dropout ensembles, LLM augmentation for a high-resource language (Russian), an existence-based decision filter, per-class thresholds, and cross-lingual transfer from the bilingual model to English.

A recurring theme is the unreliability of development-set performance as a proxy for test performance under calibration shift. The clearest instance is the inversion between our two English models: the augmented mDeBERTa-v3-base model reaches development F1 of 0.7902, which is 0.053 above BiomedBERT (0.7369), yet is outperformed on the test set by 0.140 F1 points (0.3334 vs. 0.4733). We attribute this to calibration degradation introduced by LLM-generated training data, which pushes the model toward overconfident predictions for augmented classes while degrading probability estimates elsewhere.

The remainder of the paper is organised as follows. Section 2 reviews related work; Section 3 describes the task data and the distributional properties motivating our design; Section 4 presents the system; Section 5 reports development and test results; Section 6 documents our negative findings; Section 7 discusses broader implications; and Section 8 concludes.

2. Related Work

2.1. Relation Extraction with Entity Marker Representations

The standard approach to sentence-level RE with pre-trained language models such as BERT [3] inserts special boundary tokens around the head and tail spans and pools the hidden states at those positions for classification. This *entity marker* formulation consistently outperforms alternatives such as mention pooling or entity masking on general-domain benchmarks such as TACRED [4], because the markers both signal entity boundaries to the attention mechanism and provide fixed pooling positions whose representations become task-specific during fine-tuning. Zhong and Chen [5] established a clean empirical baseline with this approach, showing that *typed* entity markers, where the marker token itself encodes the entity type, improve over type-agnostic markers, which they attribute to the model's ability to exploit entity-type constraints on the relation type. Later work extends the marker idea: Ye

[1] Our code and prediction files are available at https://anonymous.4open.science/r/NEREL-BIO-private-B9B6/.

et al. [6] pack multiple span markers into a single encoder pass and use *levitated* markers that detach marker positions from the surface text, improving accuracy and efficiency on overlapping spans, while span-based joint models such as SpERT [7] enumerate candidate spans and classify their relations without explicit boundary tokens.

A subsequent analysis qualifies how much of the typed-marker gain is genuinely relational. Mtumbuka and Schockaert [8] show that relation representations formed by concatenating head and tail entity-marker embeddings predominantly capture the *types* of the entities involved, which can cause confusion among relations that hold between entities of the same type. For a constrained label space such as BioNNE-R this property is largely a feature: entity-type pairs strongly restrict the admissible relation types (for example, CHEM → DISO is compatible with TREATED_USING, AFFECTS, and HAS_CAUSE but not with PHYSIOLOGY_OF or PART_OF), and making this prior directly visible to the encoder provides a legitimate inductive bias. Nested spans nonetheless introduce a complication that the marker literature does not address: when one span contains the other, naive left-to-right insertion of the four marker tokens corrupts the character offsets of at least two of them. We address this directly in Section 4.2.

2.2. Domain-Specific Pretraining for Biomedical NLP

Across biomedical benchmarks, models pre-trained from scratch on in-domain text outperform those adapted from general-domain checkpoints. Gu et al. [9] demonstrated this with BiomedBERT (originally PubMedBERT), pre-trained exclusively on PubMed abstracts and PubMed Central full text, which reached state-of-the-art results across the BLURB suite. Much of the advantage comes from vocabulary alignment. A domain-specific tokeniser produces fewer fragmented subword tokens for biomedical terminology, which yields cleaner entity-boundary representations and more coherent span pooling, the operation on which the marker-based classification head depends. For multilingual settings, DeBERTaV3 [10] provides a strong general-purpose encoder through disentangled attention, which encodes content and relative-position information in separate vectors. This separation is useful for RE because entity position and entity content are both informative for relation type, and encoding them independently reduces the risk of conflating positional artefacts with semantic relations. In the BioNNE-L 2025 entity-linking task on the same corpus, domain-adapted and multilingual encoders consistently outperformed general multilingual baselines [11].

2.3. The NEREL-BIO Corpus and the BioNNE Task Series

NEREL-BIO [2] is a bilingual biomedical corpus of PubMed abstracts with nested named-entity annotation. It extends the general-domain NEREL dataset [12], which annotated nested entities, relations, and events over Russian news and Wikipedia text and supplied the relation taxonomy that NEREL-BIO adapts for the biomedical domain. The BioNNE series at BioASQ provides a longitudinal evaluation on this corpus, run within the wider BioASQ challenge, including the fourteenth BioASQ lab at CLEF 2026 [13, 14, 15]. BioNNE 2024 addressed nested named-entity recognition [16]; the entity-linking task BioNNE-L 2025 reported that top systems used contrastive fine-tuning of biomedical encoders [11]; and BioNNE-R 2026 completes the pipeline by targeting relation extraction between the nested entities [17].

2.4. Data Augmentation with Large Language Models

Using large language models to synthesise training examples has become a practical strategy for low-resource and class-imbalanced NLP. Ali et al. [18] apply LLM-based augmentation to relation extraction and report improvements under low-resource conditions, with the largest gains on sparse relation classes. A generate-then-verify architecture, in which a cheaper model generates candidates and a stronger model filters them, is a natural way to control register and label correctness, both difficult to guarantee for GPT-generated biomedical text. Surveys of LLM augmentation [19] identify distributional shift between generated and human-authored text as a recurring failure mode, in which a fine-tuned model assigns atypically high confidence to generated examples and so loses calibration. In

the biomedical setting, Delmas et al. [20] show that synthetic data and diversity-optimised sampling improve relation extraction in underexplored, low-resource domains, while the benefit diminishes as the volume of real data grows. Our experiments reproduce both sides of this picture. The effect of augmentation is sharply asymmetric: at best marginal for the very small English training set, and clearly harmful for the much larger Russian one (Sections 5 and 6).

2.5. Confidence Calibration and Threshold-Based Inference

Modern deep classifiers are systematically overconfident, and the gap between confidence and accuracy widens with model size, dataset size, and fine-tuning. Guo et al. [21] provided the seminal characterisation of this phenomenon and showed that simple post-hoc temperature scaling restores calibration across many architectures. For RE under negative sampling, a specific mismatch arises: when all within-document pairs are enumerated at inference time, negatives are far more prevalent than during training, so a classifier at a fixed threshold produces too many false positives. Threshold moving, i.e. choosing a decision boundary on held-out data calibrated to the deployment class distribution, is a standard remedy that extends naturally to the multi-class setting via per-prediction confidence thresholds. In document-level RE, Zhou et al. [22] make this boundary *learnable*, replacing a global threshold with an entity-pair-dependent adaptive threshold; we instead keep a single global threshold but calibrate it post-hoc on a deployment-representative development set, which requires no change to training. Xie et al. [23] propose adaptive temperature scaling, learning instance-conditional temperatures rather than a single global scalar; for our problem the mismatch is distributional rather than instance-level, and a global threshold sweep suffices. The existence filter we evaluate, which uses $1 - p(\texttt{no\_relation})$ as the decision score, is useful when the model spreads probability across positive classes rather than concentrating it on one.

3. Task, Data, and Evaluation

3.1. Task Definition

BioNNE-R frames biomedical RE as supervised multi-class classification. Given a document and a set of pre-annotated entity mentions with character spans and types, systems predict the directed relation between each candidate pair. Candidate pairs are all within-sentence entity pairs from the provided entity annotation; systems output only pairs assigned a positive relation, omitting those classified as `no_relation`. The label space comprises fourteen directed relation types: ABBREVIATION, AFFECTS, ALTERNATIVE_NAME, APPLIED_TO, ASSOCIATED_WITH, FINDING_OF, HAS_CAUSE, ORIGINS_FROM, PART_OF, PHYSIOLOGY_OF, SUBCLASS_OF, TO_DETECT_OR_STUDY, TREATED_USING, and USED_IN.

The evaluation metric is **macro-averaged F1** over the fourteen positive classes; `no_relation` does not contribute. Every class therefore contributes equally regardless of frequency: a system that is near-perfect on the frequent classes but produces no true positives for the three rarest classes is capped well below its apparent quality. Rare-class coverage is thus a first-order concern, motivating both the augmentation strategy (Section 4.5) and the class-weighted loss (Section 4.4). The three subtasks are scored independently; the bilingual track requires a single model trained on combined data, and prediction files for any two tracks must differ.

3.2. Dataset Overview

The data is drawn from NEREL-BIO. Table 1 summarises the splits. The data is strikingly asymmetric between languages. The Russian training partition has 717 documents and 23,773 relations, while English has only 56 documents and 3,193 relations, a $7.4\times$ difference in positive training instances. This reflects the genuine scarcity of annotated Russian biomedical text rather than a design choice, and means English behaves as a low-resource scenario in practice. English biomedical RE, by contrast, is

Table 1
Dataset statistics. Test relation counts are blind. The English training set is 7.4× smaller than Russian in annotated relations.

Split	Documents	Entities	Gold relations
EN train	56	3,862	3,193
EN dev	51	3,479	2,968
EN test	154	10,317	n/a
RU train	717	37,880	23,773
RU dev	51	3,211	2,522
RU test	155	9,340	n/a

served by several multi-type corpora such as BioRED [24] and the ChemProt/BioCreative VI track [25], within a broad and well-surveyed dataset landscape [26]; no comparable Russian resource exists, which motivates the NEREL-BIO annotation effort. Eight entity types are annotated in both languages: DISO (disorders), ANATOMY, CHEM (chemicals/drugs), PHYS (physiological processes), FINDING, LABPROC (laboratory/diagnostic procedures), DEVICE, and INJURY_POISONING. DISO and ANATOMY are the most frequent.

3.3. Relation Distribution and Per-Class Difficulty

Table 2 reports the English relation distribution together with the per-class development F1 of our best submitted English model. The distribution is strongly skewed: SUBCLASS_OF is the most frequent class (743 instances), partly because nestedness makes a same-type containment (*e.g.*, *dermatitis* inside *atopic dermatitis*) a frequent surface realisation of subclass relations. APPLIED_TO has only 43 training instances and is further confused with TO_DETECT_OR_STUDY: both share the dominant LABPROC → ANATOMY type pair, distinguished by whether a procedure is interventional or observational.

ALTERNATIVE_NAME is the clearest case. With 95 training instances and 98 gold development instances, it has nearly the same training count as ABBREVIATION (97 instances), yet reaches dev F1 of only 0.2482 against ABBREVIATION's 0.8627. What separates them is surface-form regularity rather than data quantity. ABBREVIATION follows a near-universal parenthetical pattern (*hypertension (HTN)*), whereas ALTERNATIVE_NAME spans brand-vs-generic drug names, historical-vs-modern terminology, transliterated forms, and full synonym substitutions, a diversity that 95 examples cannot cover. A confusion analysis confirms the failure is one of detection rather than mislabelling: 61 of the 98 gold ALTERNATIVE_NAME instances (62%) are predicted as `no_relation`. This diagnostic motivates the targeted augmentation in Section 4.5.

3.4. Nestedness and Span Overlap

Approximately 15% of within-sentence entity pairs involve overlapping spans.[2] These nested pairs are not uniformly distributed: SUBCLASS_OF and PART_OF are predominantly nested, while TREATED_USING, AFFECTS, and HAS_CAUSE predominantly involve external (non-overlapping) pairs. A model that handles external pairs well but fails on nested ones will systematically underperform on two frequent classes, motivating the nesting-aware encoding of Section 4.2.

3.5. The Train-to-Test Negative Ratio Mismatch

During training, negative pairs are sampled at 3:1 relative to positives. At test time the task requires predicting over *all* within-sentence pairs. The English test set (154 documents, 10,317 entities) produces ≈818K candidate pairs and the Russian set ≈664K. Scaling the gold positive count from the development

[2]Computed from the NEREL-BIO training-set annotations; an approximate corpus statistic, not a figure reported by the dataset paper.

Table 2
English training-set relation distribution and per-class development F1 of the best submitted English model (BiomedBERT v2), under the blind protocol (Section 5.1); the per-class scores macro-average to the reported 0.7369. Sorted by training count; the two chronic underperformers are shown in bold.

Relation	Train	Dev gold	Dev F1
SUBCLASS_OF	743	757	0.8556
HAS_CAUSE	579	487	0.7644
AFFECTS	362	393	0.8996
ASSOCIATED_WITH	325	233	0.7485
PART_OF	204	248	0.7452
TO_DETECT_OR_STUDY	164	116	0.7181
PHYSIOLOGY_OF	150	178	0.9080
FINDING_OF	121	83	0.7432
ORIGINS_FROM	120	79	0.8024
TREATED_USING	101	88	0.8118
ABBREVIATION	97	74	0.8627
ALTERNATIVE_NAME	95	98	**0.2482**
USED_IN	89	79	0.7250
APPLIED_TO	43	54	**0.4835**
Total	**3,193**	**2,968**	n/a

set, we expect roughly 8,950 (EN) and 7,650 (RU) positive test relations, giving effective negative ratios of $\approx$91:1 and $\approx$87:1, about 30$\times$ the training ratio. A softmax classifier trained at 3:1 is therefore systematically overconfident about positives at test time: at the default threshold of 0.5 our mDeBERTa models produce 33,338 (EN) and 17,982 (RU) predictions, well above expected counts, with macro F1 of 0.2819 and 0.4084, far below their development values. As shown later in Figure 2, the threshold calibration of Section 4.6 corrects this mismatch.

4. System Description

Figure 1 summarises the system: nested-entity input and typed-marker construction, the domain-appropriate encoders and classification head, and the two-stage LLM augmentation pipeline for rare classes. The rest of this section describes each component, and inference-time calibration follows in Section 4.6.

4.1. Model Selection and Backbone Encoders

Our system follows the entity-marker RE paradigm: a pre-trained encoder processes the sentence with special tokens at entity boundaries, and a classification head over the concatenated marker representations predicts the relation. We select backbones per track rather than using one multilingual model throughout. For **English (Subtask 1)** we use `BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext` [9], pre-trained on PubMed abstracts and PubMed Central full text (the same text type as the task documents), with a domain-specific vocabulary that fragments biomedical terms far less than a general tokeniser. For **Russian and Bilingual (Subtasks 2–3)** we use `mdeberta-v3-base` [10], whose disentangled attention keeps positional and content signals separate until the attention computation, a useful property for RE. At 278M parameters mDeBERTa is larger than BiomedBERT (110M), requiring mixed precision and gradient checkpointing for the Russian training set. We also trained an mDeBERTa model on the English data and applied the bilingual model to the English test set as experimental conditions (Sections 5 and 6).

All models trained for 10 epochs with AdamW, linear decay, and 300 warmup steps; the best checkpoint was selected by blind-development macro F1 (Section 5.1). Training used a dual-checkpoint scheme for fault tolerance on time-limited cloud GPU sessions: a best-model checkpoint saved on F1

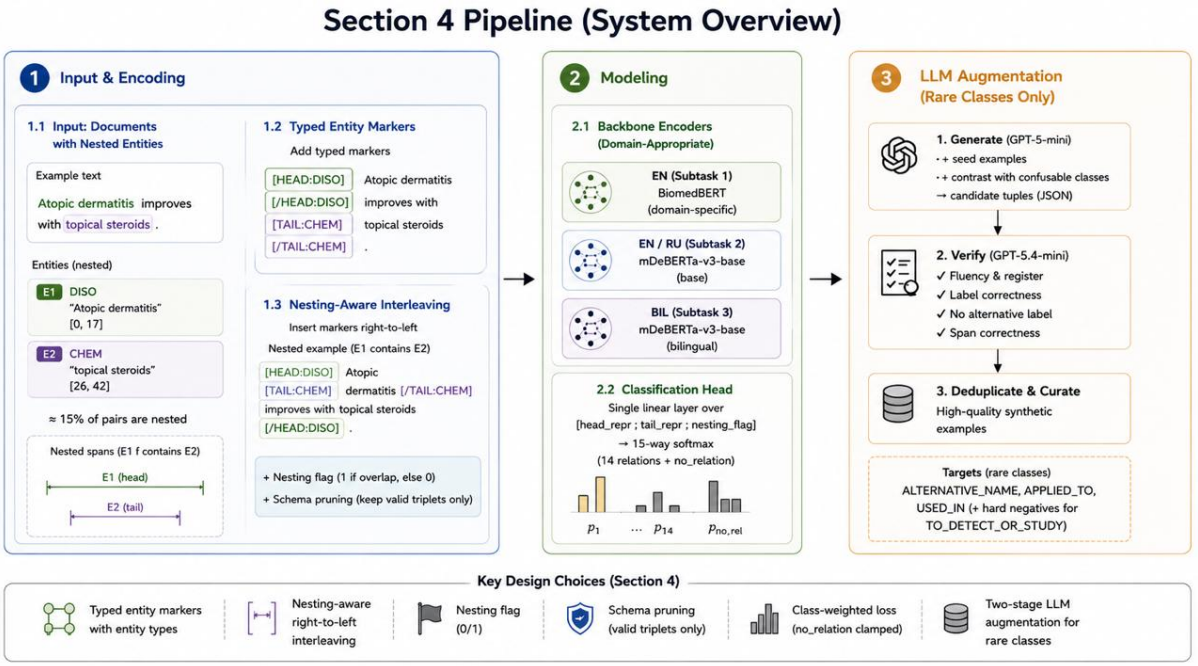

Figure 1: System overview. Panel 1: documents with nested entities are encoded with typed entity markers, inserted by nesting-aware right-to-left interleaving, together with a nesting flag and schema pruning. Panel 2: domain-appropriate encoders (BiomedBERT for English; mDeBERTa-v3-base for Russian and bilingual) feed a single linear head over the concatenated marker representations, producing a 15-way softmax (14 relations plus `no_relation`). Panel 3: the two-stage generate-then-verify LLM augmentation pipeline targets rare classes (Section 4.5).

Table 3
Training hyperparameters.

Parameter	BiomedBERT (EN)	mDeBERTa (RU/BIL)
Learning rate	2e-5	2e-5
Batch size	16	128
Max sequence length	512	512
Epochs	10	10
Warmup steps	300	300
Mixed precision	no	fp16
Gradient checkpoint	no	yes
Optimizer	AdamW, wd 0.01	AdamW, wd 0.01
LR schedule	linear decay	linear decay

improvement, and a full resume checkpoint (optimiser and scheduler state) saved every epoch and every 300 steps. Table 3 lists the hyperparameters. Unless otherwise stated, runs used random seed 42 for data shuffling, negative sampling, model initialisation, and MC-dropout sampling. Experiments were run with Python 3.11, PyTorch 2.5, Transformers 4.46, and CUDA 12 on NVIDIA A100-class GPUs; the exact hardware varied across time-limited cloud sessions, so checkpoint selection rather than wall-clock time is the reproducibility target.

4.2. Typed Entity Markers with Nesting-Aware Interleaving

The organiser baseline, built on OpenNRE [27], uses four generic special tokens (`[unused0]`–`[unused3]`). We replace them with 32 typed markers of the form `[HEAD:TYPE]`, `[/HEAD:TYPE]`, `[TAIL:TYPE]`, `[/TAIL:TYPE]`, covering all eight entity types in head and tail roles. Markers are inserted at character-level span boundaries before tokenisation, and the entity representation is the

concatenation of the hidden states at the [HEAD:TYPE] and [TAIL:TYPE] opening positions, following the start-token pooling of Zhong and Chen [5]. For the sentence *"lisinopril reduces blood pressure"* with head *lisinopril* (CHEM) and tail *blood pressure* (PHYS), the encoded form is:

```
[HEAD:CHEM] lisinopril [/HEAD:CHEM] reduces [TAIL:PHYS] blood
    pressure [/TAIL:PHYS]
```

The markers expose the entity-type pair to the encoder directly, ruling out incompatible relation types before any contextual reasoning.

Nesting-aware interleaving. When one span contains the other, for instance head *atopic dermatitis* (DISO, characters 0–17) and tail *dermatitis* (DISO, characters 7–17), naive left-to-right insertion shifts later character positions by the length of each inserted marker, corrupting the remaining offsets. We insert all markers in **right-to-left order** over the original positions: each insertion shifts only positions to its left, which have not yet been used, so every remaining insertion point stays valid. The nested example yields a well-formed, correctly interleaved sequence:

```
[HEAD:DISO] atopic [TAIL:DISO] dermatitis [/TAIL:DISO] [/HEAD:DISO]
```

Nesting flag. A binary scalar nesting flag (1.0 if the spans overlap, else 0.0) is appended to the concatenated entity representation before the head, forming an input of dimension $2H + 1$ where H is the encoder hidden size. This supplies an explicit structural signal for the $\approx 15\%$ of pairs whose relation-type distribution differs qualitatively from non-overlapping pairs.

4.3. Valid Triplet Schema Pruning

Early experiments showed the model predicting schema-invalid relations ($e.g.$, AFFECTS(ANATOMY, ANATOMY)) that have zero training instances. We extracted all valid `(head_type, relation, tail_type)` triplets from the official NEREL-BIO annotation schema configuration (the type–relation–type combinations licensed by the guidelines, not the subset observed in training), obtaining **82** valid combinations out of a theoretical $8 \times 14 \times 8 = 896$. At inference, any predicted relation violating this schema is overridden to `no_relation` before thresholding. The filter is computationally free and removed hallucinated positives without affecting valid predictions.

4.4. Class-Weighted Loss with `no_relation` Clamping

Per-class loss weights are the inverse of training frequency, normalised so the mean positive-class weight is 1.0; rare classes such as APPLIED_TO and ALTERNATIVE_NAME receive roughly 4–$8\times$ the weight of frequent classes. A naive scheme would also assign `no_relation` a small weight reflecting its high training frequency, but `no_relation` is *artificially under-represented* at training time (3:1) relative to test ($\approx$90:1), so a low weight would make negative-class false negatives costly, the opposite of what is needed. We therefore **clamp the `no_relation` weight to 1.0** and apply inverse-frequency weighting only to the fourteen positive classes.

4.5. Two-Stage LLM Augmentation for Rare Classes

The augmentation targets two English classes identified by the confusion analysis. ALTERNATIVE_NAME (95 train, dev F1 0.2482) fails by detection: 62% of gold development instances are predicted as `no_relation`, a consequence of surface-form diversity rather than data quantity. APPLIED_TO (43 train, dev F1 0.4835) fails by confusion: 28% of gold development instances are labelled TO_DETECT_OR_STUDY, the two relations sharing the LABPROC $\rightarrow$ ANATOMY pair and differing only in interventional vs. observational intent. For Russian we additionally targeted USED_IN.

Table 4
LLM augmentation targets and accepted counts. Accepted counts are approximate where verification rejected part of a batch.

Lang	Relation	Type pair	Target	Accepted
EN	ALTERNATIVE_NAME	DISO–DISO	120	≈120
EN	ALTERNATIVE_NAME	CHEM–CHEM	90	≈90
EN	ALTERNATIVE_NAME	ANATOMY–ANATOMY	80	≈80
EN	ALTERNATIVE_NAME	FINDING–FINDING	60	≈53
EN	ALTERNATIVE_NAME	LABPROC–LABPROC	30	≈25
EN	ALTERNATIVE_NAME	PHYS–PHYS	20	≈20
EN	APPLIED_TO	LABPROC–ANATOMY	130	≈110
EN	APPLIED_TO	CHEM–ANATOMY	70	≈60
EN	TO_DETECT_OR_STUDY	LABPROC–DISO,CHEM	150	≈131
RU	ALTERNATIVE_NAME	(multiple)	320	≈326
RU	APPLIED_TO	(multiple)	200	≈193
RU	TO_DETECT_OR_STUDY	LABPROC–ANATOMY,PHYS	130	≈187
RU	USED_IN	(multiple)	160	≈168
Total EN				**917**
Total RU				**874**

Generation (GPT-5-mini). For each target relation and entity-type pair we prompt the generator with (i) a precise relation definition, (ii) five seed examples sampled from gold training data matching the type pair, (iii) an explicit contrast with the most confusable neighbour, and (iv) a request for five new tuples `(sentence, head, head_span, tail, tail_span, relation)` in JSON, varying sub-domain and syntax. For APPLIED_TO jobs we additionally request five *hard-negative* TO_DETECT_OR_STUDY examples with the same type pair but observational framing, included in training under the TO_DETECT_OR_STUDY label to teach the decision boundary directly. Malformed JSON is discarded without retry. Generation used temperature 0.7, nucleus sampling with p=0.95, no frequency or presence penalty, and a 4096-token output cap.

Verification (GPT-5.4-mini). Each candidate is independently checked for (i) fluency and publication register, (ii) unambiguous labellability from sentence context alone, (iii) the absence of an equally valid alternative label, and (iv) correct character spans; only candidates accepted on all four criteria are kept. Before the verifier is called, programmatic span validation requires `sentence[head_start:head_end]` to match the head string within a $\pm$3-character tolerance, discarding unlocatable candidates and saving cost. Final counts are 917 English examples (ALTERNATIVE_NAME 413, TO_DETECT_OR_STUDY hard negatives 301, APPLIED_TO 203) and 874 Russian (ALTERNATIVE_NAME 326, APPLIED_TO 193, TO_DETECT_OR_STUDY 187, USED_IN 168); Table 4 gives the breakdown. One Russian APPLIED_TO(LABPROC, ANATOMY) job repeatedly hit the output-token limit because Russian procedure names are morphologically longer, reaching 86% of its target. Verification used deterministic decoding (temperature 0) with the same output cap. The generation and verification prompt templates are given in Appendix B.

4.6. Blind-Development Confidence Threshold Calibration

As quantified in Section 3.5, the test negative ratio is about 30$\times$ the training ratio, so a model at t=0.5 over-predicts positives. We run a single inference pass over all candidate pairs from development documents, at the same ≈87:1 ratio as the test set, and sweep thresholds $T \in [0, 1]$ over the stored softmax tensors, computing predicted counts and macro F1 against gold annotations in $O(N)$ CPU time; the T maximising blind-development macro F1 is used for submission. We additionally evaluate an **existence filter** that uses $1 - p(\texttt{no\_relation})$ as the decision score, decoupling whether *any*

relation exists from which type it is, motivated by the possibility that for miscalibrated models the positive probability mass is spread across several class logits.

4.7. Truncation Failsafe

At `max_length=512`, about 3% of instances, typically long sentences with distant pairs, have one or both markers fall outside the window after truncation, causing the model to pool an arbitrary token as the entity representation. These errors are silent: the input tensor is well formed and the output is confident but meaningless. Our first English model (v1), trained without this protection, reached only 0.2524 blind-development F1. Tracking marker validity at dataset construction and forcing `no_relation` when a marker is absent resolved the failure; the corrected v2 model reached 0.7369 (Section 5.2).

5. Results

This section is organised around the questions a reader is likely to ask. Section 5.1 fixes the evaluation protocol, which differs from the organiser baseline and is needed to read every number that follows. Section 5.2 reports an instructive engineering failure and the one-line fix that resolved it. Section 5.3 gives the best development score per track, and Section 5.4 isolates the effect of threshold calibration. Section 5.5 presents the full submission history and the final test scores. The last two subsections analyse the findings behind the paper's title: the development-to-test inversion (Section 5.6) and the rare-class coverage that explains it (Section 5.7).

5.1. Development Evaluation Protocol

Development scores are reported in *blind mode*: inference runs over all candidate pairs from development documents (the same $\approx$87:1 ratio as the test set), then macro F1 is computed against gold annotations. This contrasts with evaluating only the labelled subset of pairs, which inflates scores by excluding most negatives. Comparisons with the organiser baseline, which uses the labelled-subset protocol, should be read with this difference in mind.

5.2. The BiomedBERT v1 Failure

The first English model reached blind-development F1 of **0.2524**, below the organiser baseline. The cause was the truncation failure described in Section 4.7: silent marker loss let the classifier adapt to a meaningless signal. Applying the truncation failsafe, the only change between the two runs, raised the score to **0.7369**. We report this as a cautionary finding: silent input corruption is invisible to the training loss and can masquerade as a modelling problem.

5.3. Best Development Results

Table 5 reports the best blind-development F1 per track. The augmented mDeBERTa English model is highest overall (0.7902), 0.053 above BiomedBERT. Critically, this advantage does not transfer to the test set (Section 5.6). Its learning curve is also revealing: development F1 peaks sharply at epoch 2 and never recovers, oscillating between 0.7206 and 0.7759 thereafter (Table 6), consistent with rapid memorisation of the augmented examples followed by failure to generalise. The Russian and bilingual models, trained on larger data without augmentation, instead converge monotonically by epoch 7–8.

5.4. Threshold Calibration Effect

At t=0.5 the mDeBERTa models far exceed expected prediction counts: 33,338 (EN, F1 0.2819), 17,982 (RU, F1 0.4084), and 58,486 (BIL, F1 0.3310). The blind-development sweep selects T=0.95 for Russian and Bilingual and T=0.80 for BiomedBERT English. Calibration is the largest single intervention

Table 5
Best blind-development macro F1 per track.

Track	Model	Dev macro F1
EN	BiomedBERT + typed markers	0.7369
EN	mDeBERTa-v3-base + augmentation	**0.7902**
RU	mDeBERTa-v3-base (base)	**0.8397**
BIL	mDeBERTa-v3-base (bilingual)	**0.8831**

Table 6
Per-epoch blind-development F1 of the augmented English mDeBERTa model (rounded to three decimals; the epoch-2 peak is 0.7902). The best checkpoint (epoch 2) is retained.

Epoch	1	2	3	4	5	6	7	8	9	10
F1	0.634	**0.790**	0.726	0.776	0.721	0.766	0.755	0.770	0.762	0.754

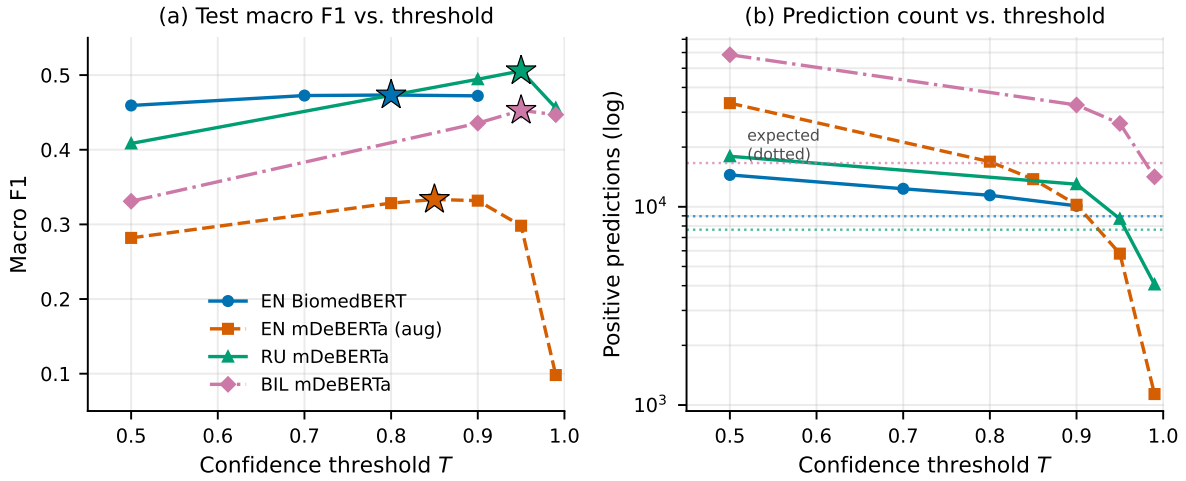

Figure 2: Threshold calibration on the test set, for the best model on each track. (a) Macro F1 versus confidence threshold T; stars mark the threshold selected on blind development. (b) Positive-prediction count versus T (log scale); dotted lines give the expected positive counts per track. At T=0.5 every model over-predicts and scores poorly; the selected thresholds bring prediction counts close to expectation and maximise macro F1. Operating points are the submitted thresholds (see text).

across all tracks: for Russian, moving from t=0.5 to t=0.95 raises macro F1 from 0.4084 to **0.5057** (+0.097) with no change to weights, architecture, or data; for Bilingual the corresponding gain is +0.122. Figure 2 traces both effects. Because we did not persist the per-pair probability tensors beyond the time-limited cloud GPU sessions used for inference, the curves connect the thresholds we actually submitted rather than a continuous sweep; the trend is nonetheless unambiguous. The figure also previews the calibration contrast of Section 5.6: BiomedBERT's English curve is almost flat near its optimum, the signature of a well-calibrated model, whereas the augmented English mDeBERTa peaks sharply at t=0.85 and then collapses.

5.5. Submission History and Test Results

Tables 7 and 8 give the full submission search. Three patterns recur across them. First, in every track the gain from calibration exceeds that from any single modelling choice. Second, the augmented English mDeBERTa underperforms BiomedBERT at every threshold, confirming that its development advantage does not reflect better generalisation. Third, the bilingual model applied to English peaks at 0.4542, below BiomedBERT's 0.4733, despite more training data. Final scores are summarised in Table 9.

Table 7
Complete English (Subtask 1) submission history, sorted by macro F1. The final best submission is in bold.

Model	Threshold / filter	Predictions	Macro F1
BiomedBERT	**0.80**	**11,430**	**0.4733**
BiomedBERT	0.70	12,329	0.4726
BiomedBERT	0.90	10,113	0.4722
BiomedBERT	0.50	14,480	0.4593
mDeBERTa BIL on EN	0.95	14,328	0.4542
mDeBERTa BIL on EN	0.90	17,877	0.4447
mDeBERTa BIL on EN	0.99	7,657	0.4400
mDeBERTa EN aug	existence 0.95	16,349	0.3870
mDeBERTa EN aug	per-class	10,616	0.3625
mDeBERTa EN aug	0.85	13,785	0.3334
mDeBERTa EN aug	0.90	10,239	0.3318
mDeBERTa EN aug	0.95	5,804	0.2982
mDeBERTa EN aug	0.50	33,338	0.2819
mDeBERTa EN aug	0.99	1,136	0.0982

Table 8
Russian (Subtask 2) and Bilingual (Subtask 3) submission history. Best per track in bold.

Track	Model	Threshold	Predictions	Macro F1
RU	**mDeBERTa base**	**0.95**	**8,692**	**0.5057**
RU	mDeBERTa base	existence 0.95	10,418	0.5036
RU	mDeBERTa base	0.90	$\approx$13,000	0.4944
RU	mDeBERTa base	0.99	4,057	0.4560
RU	mDeBERTa base	0.50	17,982	0.4084
RU	mDeBERTa aug	0.50	$\approx$33,000	0.2990
BIL	**mDeBERTa BIL**	**0.95**	**26,318**	**0.4527**
BIL	mDeBERTa BIL	existence 0.95	32,434	0.4505
BIL	mDeBERTa BIL	0.99	14,163	0.4470
BIL	mDeBERTa BIL	0.90	32,618	0.4356
BIL	mDeBERTa BIL	0.50	58,486	0.3310

Table 9
Final best scores per track.

Track	Macro F1	Model	T	Predictions
EN (Subtask 1)	**0.4733**	BiomedBERT	0.80	11,430
RU (Subtask 2)	**0.5057**	mDeBERTa-v3 (base)	0.95	8,692
BIL (Subtask 3)	**0.4527**	mDeBERTa-v3 (bilingual)	0.95	26,318

5.6. The Development-to-Test Inversion

On English, the development and test rankings invert (Figure 3). BiomedBERT scores $0.7369 \rightarrow 0.4733$ while the augmented mDeBERTa scores $0.7902 \rightarrow 0.3334$. The model with the lower development F1 wins on test by 0.140. Three factors explain this.

Calibration. BiomedBERT's confidence distribution is smooth and unimodal: macro F1 varies by less than 0.001 across $T \in [0.70, 0.90]$. The augmented mDeBERTa's distribution is bimodal, with a gap in $[0.90, 0.99]$: prediction count collapses from 10,239 at t=0.90 to 1,136 at t=0.99, a $9\times$ drop over a 0.09 change, because GPT-generated examples are stylistically more regular than real PubMed text, drawing very high confidence for instances that resemble them. *Vocabulary match.* BiomedBERT tokenises novel biomedical terms faithfully, yielding reliable marker representations; mDeBERTa fragments them. *Augmentation overfitting.* The 917-example set is too small to regularise a 278M-parameter model over

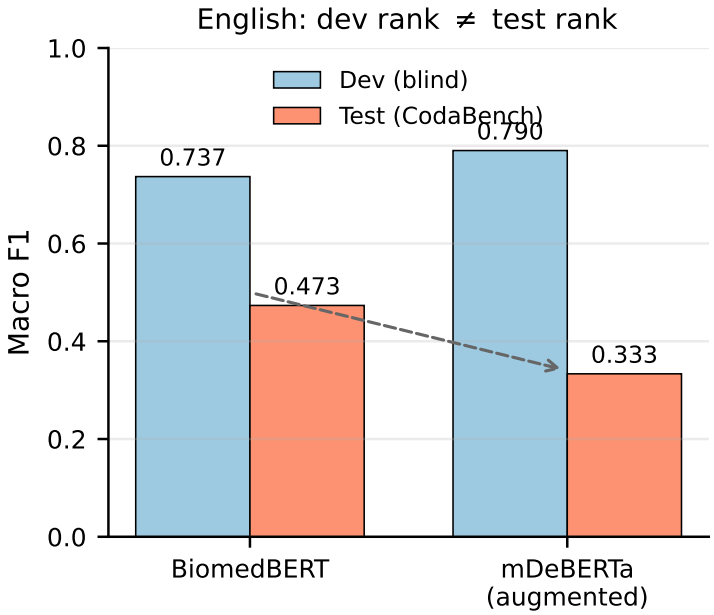

Figure 3: The English development-to-test inversion. The augmented mDeBERTa model has the higher blind-development F1 but the lower test F1; BiomedBERT, with a lower development score, generalises better. Development rank does not predict test rank under calibration shift.

10 epochs, producing the epoch-2 peak and subsequent oscillation.

5.7. Rare-Class Coverage

Table 10 compares per-class predicted counts for English, computed directly from the submitted prediction files. Despite targeting ALTERNATIVE_NAME, APPLIED_TO, and USED_IN, the augmented mDeBERTa nearly zeroes all three at the threshold that suppresses the over-predicted common classes: at t=0.85 it emits 12, 3, and 41 predictions respectively, against expected counts of ≈296, ≈163, and ≈239. BiomedBERT, at its best threshold (t=0.80), predicts 32, 31, and 241, an order of magnitude more APPLIED_TO and roughly six times more USED_IN, covering the latter almost exactly (101% of expected). Because macro F1 weights all classes equally, the augmented model's near-zero coverage on three classes caps it far below BiomedBERT. ALTERNATIVE_NAME remains poorly covered by both models, a difficulty that neither domain pretraining nor augmentation resolves at the available data volume. The bilingual model, applied to English, recovers USED_IN coverage (599) through cross-lingual transfer but over-predicts ORIGINS_FROM (1,003 vs. ≈239 expected), importing Russian-specific associations (Section 6.5).

6. Negative Results

The experiments below were deliberate hypothesis tests, not post-hoc rationalisations. Each follows the same structure (hypothesis, experiment, result, interpretation) because negative results from well-resourced shared-task systems are underrepresented, and each carries actionable guidance.

6.1. MC-Dropout Stochastic Ensembles

Hypothesis. Monte Carlo dropout [28], i.e. multiple stochastic forward passes with dropout enabled at inference and aggregated by plurality vote, reduces variance and improves macro F1. **Experiment.** We combined one deterministic pass with three stochastic passes (different seeds) on the organiser baseline and on our fine-tuned models, voting per instance with ties broken toward `no_relation`. **Result.** The

Table 10
English test predicted relation counts, computed from the submitted prediction files, at each model's best threshold. Expected counts are the gold development counts scaled by the test/development document ratio ($154/51 \approx 3.02$). The BIL $\rightarrow$ EN column is the bilingual model applied to the English test set (Section 6.5). The three augmentation-target classes are in bold.

Class	Exp. (≈)	BiomedBERT t=0.80	mDeBERTa aug t=0.85	BIL $\rightarrow$ EN t=0.95
SUBCLASS_OF	2,286	3,096	2,451	2,493
HAS_CAUSE	1,471	1,391	2,498	1,738
AFFECTS	1,187	1,803	2,145	1,470
PART_OF	749	1,021	715	1,082
ASSOCIATED_WITH	704	1,263	1,619	1,983
PHYSIOLOGY_OF	537	640	1,366	1,260
TO_DETECT_OR_STUDY	350	548	782	966
TREATED_USING	266	250	418	303
FINDING_OF	251	489	853	1,048
ORIGINS_FROM	239	323	354	1,003
ABBREVIATION	223	302	528	275
USED_IN	239	**241**	**41**	599
ALTERNATIVE_NAME	296	**32**	**12**	76
APPLIED_TO	163	**31**	**3**	32
Total	8,959	11,430	13,785	14,328

ensemble produced no measurable improvement; for the baseline, development F1 stayed within 0.002 of deterministic inference across all tracks. It was not adopted. **Interpretation.** MC dropout helps under high epistemic uncertainty. After supervised fine-tuning on domain-matched representations, the posterior collapses tightly around the fine-tuned weights, leaving little uncertainty for the ensemble to resolve, consistent with the framing of Gal and Ghahramani [28].

6.2. LLM Augmentation is Net Negative for Russian

Hypothesis. The pipeline that helps English rare classes will help Russian symmetrically. **Experiment.** We trained an augmented Russian mDeBERTa on the full Russian set plus 874 generated examples, with identical hyperparameters and loss. **Result.** The augmented model failed at every threshold: 0.2990 at t=0.5 versus the base model's 0.5057 at t=0.95. At t=0.99 it produced only 2,214 predictions with a pathological distribution (8 HAS_CAUSE predictions versus ≈1,030 in the base model). It did not improve a single class. **Interpretation.** The 874 examples are only 3.7% of the 23,773 Russian positives, too few to add signal yet enough to introduce distributional shift from the LLM's register, producing the same calibration collapse seen for English but more severe. A density threshold emerges: augmentation helps when generated examples are a large fraction of class-specific data (English APPLIED_TO: 203 added to 43 base, $+472\%$) and harms when they are a minor perturbation of a large set, echoing the diminishing returns reported by Delmas et al. [20].

6.3. The Existence Filter Helps Only Miscalibrated Models

Hypothesis. For models that spread positive mass across logits, $1 - p(\texttt{no\_relation})$ recovers predictions lost under a maximum-class threshold. **Experiment.** We compared the standard threshold and the existence filter at matched thresholds on all tracks via the CodaBench evaluation. **Result.** The filter improved the miscalibrated augmented English model by $+0.089$ (0.2982 $\rightarrow$ 0.3870 at t=0.95) but cost -0.002 for both the well-calibrated Russian and bilingual base models. **Interpretation.** The filter helps exactly when positive mass is diffuse. For a well-calibrated model, $\max_{c \neq \texttt{no\_relation}} p(c)$ and $1 - p(\texttt{no\_relation})$ rank instances almost identically. The size of the gain is thus a useful diagnostic for augmentation-induced calibration shift.

6.4. Per-Class Confidence Thresholds

Hypothesis. Lowering thresholds for under-predicted rare classes and raising them for over-predicted common classes will improve macro F1 over a uniform threshold. **Experiment.** On the augmented English model we set $T{=}0.40$ for ALTERNATIVE_NAME, USED_IN, APPLIED_TO; $T{=}0.95$ for PHYSIOLOGY_OF, FINDING_OF, ASSOCIATED_WITH; and $T{=}0.85$ otherwise. **Result.** Macro F1 was 0.3625, above the uniform 0.3334 but far below BiomedBERT (0.4733). At $t{=}0.40$, USED_IN over-predicted (681 vs. $\approx$239 expected) while ALTERNATIVE_NAME stayed under-predicted. **Interpretation.** Lowering the threshold for a class the model has not learned produces false positives from near neighbours, not recovered true positives. Per-class thresholds can rescue an underconfident but learned class; they cannot compensate for a representation failure caused by augmentation.

6.5. Cross-Lingual Transfer from the Bilingual Model to English

Hypothesis. The bilingual model, regularised by the larger Russian set, will transfer to English and improve rare-class coverage. **Experiment.** We applied the bilingual mDeBERTa directly to the English test set with matched threshold sweeps. **Result.** It peaked at 0.4542 ($t{=}0.95$), consistently 0.019 below BiomedBERT. Transfer did improve USED_IN coverage (599 vs. 41 for the augmented English model), but not enough to overcome BiomedBERT's advantage across the other classes. **Interpretation.** Cross-lingual transfer helps, but only so far. The Russian data supplies useful signal for semantically shared classes, yet cannot offset mDeBERTa's subword fragmentation of English biomedical terms, consistent with the BioNNE-L 2025 finding that domain-specific encoders beat multilingual ones on the English track even with more data [11].

7. Discussion

Our results across three tracks converge on three transferable principles.

7.1. Domain Pretraining Dominates Multilingual Breadth for Single-Language Tracks

As established in Section 5.6, BiomedBERT outperforms the augmented mDeBERTa on the English test set by 0.140 despite a lower development score. This raw gap, however, conflates three differences at once (domain pretraining, multilinguality, and augmentation) and therefore overstates the domain-pretraining effect in isolation: our *non-augmented* bilingual model, applied to English (Section 6.5), trails BiomedBERT by only 0.019, which suggests that much of the 0.140 is augmentation-induced calibration damage rather than domain mismatch. The residual, better-controlled advantage we attribute to two compounding factors: tokeniser alignment and pretraining-distribution match. A domain vocabulary represents terms such as *carbamazepine* or *anastomosis* as one or two tokens, concentrating meaning at the marker position the head pools, whereas a multilingual tokeniser fragments them. And BiomedBERT's PubMed pretraining already encodes the relational semantics of phrases like *used to detect*. For single-language biomedical RE, then, a smaller domain-specific encoder should be preferred to a larger general one, and a small development set may not reveal this, because it cannot expose the calibration difference that decides the test outcome.

7.2. LLM Augmentation and a Tentative Data-Density Threshold

The English-vs-Russian divergence is unlikely to be mere noise: it admits a plausible mechanism. English APPLIED_TO (43 base, $+203$, $+472\%$) and ALTERNATIVE_NAME (95 base, $+413$, $+435\%$) are dominated by the generated distribution, which genuinely broadens surface coverage. The Russian additions are a minor perturbation of a large set, so the LLM's stylistic shift outweighs the coverage gain and degrades calibration. From this single within-task comparison we can offer only a *tentative* heuristic, not a calibrated rule: LLM augmentation is more likely to help when the generated examples

at least match a class's existing training count, and more likely to harm when they are a small fraction of an already-large set; for morphologically richer languages a higher ratio may be needed. With one model on each side of a single English–Russian comparison and no statistical test, this is a hypothesis requiring validation across tasks and languages before it could be treated as a guideline. Its *direction* is at least consistent with the diminishing returns reported by Delmas et al. [20] and the distributional-shift failure mode catalogued by Ding et al. [19].

7.3. Test-Time Calibration is the Highest-Impact Intervention After Model Selection

Across all tracks the gain from moving to the blind-development optimal threshold exceeds that of any single modelling choice (e.g. +0.097 for Russian, +0.122 for Bilingual), yet it needs no extra training and no data beyond the development set. The underlying cause, the $30\times$ train-to-test negative-ratio mismatch from exhaustive pair enumeration, is structural to the RE shared-task formulation rather than specific to this dataset. We therefore argue that blind-development threshold calibration should be a standard inference component for enumerated-pair RE, analogous to how temperature scaling became standard post-hoc calibration [21, 23]. The existence-filter gain offers a companion diagnostic: large gains signal calibration damage, near-zero gains signal a well-calibrated model.

7.4. Limitations

Several limitations qualify these conclusions. First, the English test scores sit well below development, indicating overfitting to development conditions during model selection; the 56-document development set is too small to estimate macro F1 over 14 classes reliably, and a pooled or cross-validated protocol would be more robust. Second, the augmentation pipeline uses commercial API models (GPT-5-mini and GPT-5.4-mini) whose exact behaviour is not reproducible; a locally hosted verifier would make the pipeline self-contained. Third, our bilingual model trails both monolingual submissions; the task's single-model rule rightly rules out simply ensembling the two monolingual systems, so closing this gap *within* one model, for example via language routing or a mixture of experts, is an open problem we find genuinely interesting rather than a constraint to be excused. Finally, we report point estimates without confidence intervals or significance tests, and our central model contrasts are not factorially controlled (BiomedBERT differs from the augmented mDeBERTa in domain, multilinguality, and augmentation at once); the small score differences used to select operating thresholds, in particular, should be read as within sampling noise, and several thresholds were chosen with feedback from repeated leaderboard submissions rather than by the blind-development sweep alone.

8. Conclusion

We described a system for BioNNE-R 2026 reaching macro F1 of 0.4733 (English), 0.5057 (Russian), and 0.4527 (Bilingual) on the blind test set, combining typed entity markers with nesting-aware interleaving, a two-stage LLM generate-then-verify augmentation pipeline, and blind-development threshold calibration, over domain-appropriate encoders. What mattered most was the development-to-test inversion for English, where the model with the higher development score underperformed by 0.140 F1 on test. We trace this to calibration degradation from LLM-generated data and to subword fragmentation of biomedical terms. We documented five negative findings (MC dropout, Russian augmentation harm, the existence filter, per-class thresholds, and cross-lingual transfer) that together map the boundary conditions of four widely used techniques in a resource-limited biomedical RE setting. Three guidelines follow: prefer the most domain-specific available encoder for single-language tracks; treat LLM augmentation as promising only when synthetic examples can at least match a rare class's real training count (a tentative heuristic, not a rule); and treat threshold calibration on a deployment-representative development set as a mandatory inference step, since on this task it was the single largest improvement. Future work includes end-to-end joint entity-and-relation extraction, contrastive class-prototype objectives for rare-class recall, and temperature scaling calibrated to deployment-time negative ratios.

Acknowledgments

We thank the organisers of the BioNNE-R shared task and the wider BioASQ challenge for preparing the data, running the evaluation, and supporting participants.

Declaration on Generative AI

The author used generative AI tools in two clearly separated roles. As a documented part of the methodology (Section 4.5), GPT-5-mini and GPT-5.4-mini generated and verified synthetic training examples for rare relation classes; every such example passed programmatic span validation, and its effect on results is reported in full, including the cases where augmentation was harmful. In preparing the manuscript, the author additionally used a large-language-model assistant for grammar and spelling checking, paraphrasing and rewording for clarity and concision, and assistance with LaTeX formatting and the drafting of figures. Generative AI was not used to produce scientific claims, to perform the experiments, or to compute the reported numbers, all of which derive from the author's own runs. After using these tools, the author reviewed and edited all content and takes full responsibility for the publication's content.

References

[1] Y. Wang, H. Tong, Z. Zhu, Y. Li, Nested named entity recognition: A survey, ACM Transactions on Knowledge Discovery from Data 16 (2022) 1–29. doi:10.1145/3522593.

[2] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities, Bioinformatics 39 (2023) btad161. doi:10.1093/bioinformatics/btad161.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[4] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware attention and supervised data improve slot filling, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 35–45. doi:10.18653/v1/D17-1004.

[5] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2021, pp. 50–61. doi:10.18653/v1/2021.naacl-main.5.

[6] D. Ye, Y. Lin, P. Li, M. Sun, Packed levitated marker for entity and relation extraction, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers, 2022, pp. 4904–4917. doi:10.18653/v1/2022.acl-long.337.

[7] M. Eberts, A. Ulges, Span-based joint entity and relation extraction with transformer pre-training, in: Proceedings of the 24th European Conference on Artificial Intelligence (ECAI), 2020, pp. 2006–2013. doi:10.3233/FAIA200321.

[8] F. M. Mtumbuka, S. Schockaert, Entity or relation embeddings? an analysis of encoding strategies for relation extraction, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 6003–6022.

[9] 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, ACM Transactions on Computing for Healthcare 3 (2021) 1–23. doi:10.1145/3458754.

[10] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations (ICLR), 2021. ArXiv:2111.09543.

[11] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-L task on biomedical nested named entity linking in CLEF 2025, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4038/paper_3.pdf.

[12] N. Loukachevitch, E. Artemova, T. Batura, P. Braslavski, I. Denisov, V. Ivanov, S. Manandhar, A. Pugachev, E. Tutubalina, NEREL: A russian dataset with nested named entities, relations and events, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 876–885.

[13] A. Nentidis, G. Katsimpras, A. Krithara, E. Farre-Maduell, S. Lima-Lopez, M. Krallinger, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024, pp. 3–27. doi:10.1007/978-3-031-71908-0_1.

[14] A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, et al., Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025, pp. 173–198. doi:10.1007/978-3-032-04354-2_12.

[15] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-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, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[16] V. Davydova, N. Loukachevitch, E. Tutubalina, Overview of BioNNE task on biomedical nested named entity recognition at BioASQ 2024, in: CLEF 2024 Working Notes, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 28–34. URL: https://ceur-ws.org/Vol-3740/paper-03.pdf.

[17] I. Rozhkov, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[18] M. Ali, M. S. Nisar, M. Saleem, D. Moussallem, A.-C. N. Ngomo, Enhancing relation extraction through augmented data: Large language models unleashed, in: Natural Language Processing and Information Systems (NLDB 2024), volume 14763 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 68–78. doi:10.1007/978-3-031-70242-6_7.

[19] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, S. Joty, Data augmentation using large language models: Data perspectives, learning paradigms and challenges, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 1679–1705.

[20] M. Delmas, M. Wysocka, A. Freitas, Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach, Computational Linguistics 50 (2024) 953–1000. doi:10.1162/coli_a_00520.

[21] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.

[22] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 35, 2021, pp. 14612–14620. doi:10.1609/aaai.v35i16.17717.

[23] J. Xie, A. S. Chen, Y. Lee, E. Mitchell, C. Finn, Calibrating language models with adaptive temperature scaling, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

[24] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction

dataset, Briefings in Bioinformatics 23 (2022) bbac282. doi:10.1093/bib/bbac282.
[25] M. Krallinger, O. Rabal, S. A. Akhondi, et al., Overview of the BioCreative VI chemical–protein interaction track, in: Proceedings of the BioCreative VI Workshop, 2017, pp. 141–171.
[26] M.-S. Huang, J.-C. Han, P.-Y. Lin, Y.-T. You, R. T.-H. Tsai, W.-L. Hsu, Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource, Briefings in Bioinformatics 25 (2024) bbae132. doi:10.1093/bib/bbae132.
[27] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, M. Sun, OpenNRE: An open and extensible toolkit for neural relation extraction, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, 2019, pp. 169–174. doi:10.18653/v1/D19-3029.
[28] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.

A. Data Examples

Figure 4 shows representative entity mentions of each of the eight types in both languages, drawn from the NEREL-BIO training set, alongside three relation instances rendered in the typed-marker encoding of Section 4.2. The nested same-type pair is the DISO–DISO example; the TREATED_USING example is not nested.

Entity mentions by type (English / Russian)

Type	English example	Russian example (gloss)
DISO	epileptic seizure	острая респираторная инфекция (acute respiratory infection)
ANATOMY	coronary artery	носа ((of the) nose)
CHEM	carbamazepine	витамина А ((of) vitamin A)
PHYS	cardiac contractility	детская смертность (child mortality)
FINDING	seizure aggravation	заболеваемость гриппом (influenza incidence)
LABPROC	electroencephalogram	полимеразная цепная реакция (polymerase chain reaction)
DEVICE	polypropylene prosthesis	дозатор инсулина (insulin dispenser)
INJURY_POISONING	neck trauma	передозировка наркотиков (drug overdose)

Annotated relations with typed entity markers

EN, SUBCLASS_OF (nested spans):

`[HEAD:DISO] atopic [TAIL:DISO] dermatitis [/TAIL:DISO] [/HEAD:DISO]`

the inner DISO 'dermatitis' is contained in the outer DISO 'atopic dermatitis'

EN, HAS_CAUSE (external pair):

`[TAIL:CHEM] CBZ [/TAIL:CHEM] ... [HEAD:FINDING] aggravated absences [/HEAD:FINDING]`

the finding 'aggravated absences' is caused by the chemical CBZ

RU, ALTERNATIVE_NAME (same-type pair):

`[HEAD:CHEM] кофеиновая [/HEAD:CHEM] ... [TAIL:CHEM] кофеином [/TAIL:CHEM]`

two CHEM mentions of caffeine ('caffeine', adj. / instr. case)

Figure 4: Representative NEREL-BIO data. Top: one real entity mention per type for English and Russian, with the Russian glossed in English. Bottom: three relation instances shown in the typed entity-marker encoding of Section 4.2. Only the DISO–DISO SUBCLASS_OF example is nested; the TREATED_USING example with topical steroids and atopic dermatitis is a non-nested relation.

B. LLM Augmentation Prompts

The generator prompt template was:

```
You are creating biomedical relation-extraction training data.
Relation: <RELATION>. Definition: <DEFINITION>.
Head type: <HEAD_TYPE>. Tail type: <TAIL_TYPE>.
Confusable relation to avoid: <CONFUSABLE_RELATION>; contrast: <
    CONTRAST>.
Gold examples: <FIVE_GOLD_EXAMPLES>.
Generate five new PubMed-style sentences from varied biomedical
    subdomains.
Return only JSON records with sentence, head, head_span, tail,
    tail_span,
and relation. Spans are zero-based character offsets [start, end).
```

For APPLIED_TO, the generator also received: "Generate five hard-negative TO_DETECT_OR_STUDY examples with the same type pair but observational rather than interventional framing." The verifier prompt template was:

```
You are verifying one biomedical relation-extraction example.
Given the sentence, entity strings, character spans, entity types,
    and proposed
relation, answer JSON only: {"accept": true/false, "reason": "..."}.
Accept only if the sentence is fluent biomedical prose, the proposed
     relation is
entailed from the sentence alone, no alternative relation is equally
     valid, and
the character spans identify the exact entity mentions.
```