

Calibrated Typed Decisions for Entity, Relation, and Event Extraction

Yang Wang^π, Hanlin Xiao^π, Yanan Ma^κ, Chenghao Xiao^ζ, Chenghua Lin^π✉

^πUniversity of Manchester, ^κMayo Clinic, ^ζShanghai University of Finance and Economics

Abstract

Information extraction models often perform well on the datasets they are trained on, but transferring to new domains and annotation guidelines remains challenging. Supporting user-specified types alone does not resolve this problem. We present CALON, a unified 0.6B-parameter encoder for entity, relation, and event extraction across datasets and schemas. The model jointly encodes text and type definitions, then directly identifies entity spans, relations between spans, and event triggers and arguments. Its extraction decisions consider the requested types together, using their combined evidence to identify relevant candidates before assigning a type. This gives all three tasks a common probabilistic formulation while keeping every extracted mention anchored to the input text. We train on diverse sources, and supplement supervised learning with a reward that encourages correct extractions and penalises incorrect ones. Across 40 benchmarks, CALON outperforms the evaluated baselines on most benchmarks. It leads all reported relation extraction benchmarks and achieves the highest average F1 for event triggers and arguments with gold triggers. The same model also performs strongly on aspect sentiment triplet extraction and pronoun resolution through its span-pair formulation. A correction for the probability bias introduced during training improves confidence estimates for entities, relations, and event arguments after calibration, with little change in F1.

1 Introduction

Information extraction (IE) converts plain text into structured records of entities, relations, and events. The required information varies across applications, so an extractor must follow a schema specifying the types to identify. For example, job-description analysis requires identifying skills, qualifications, and experience (Green et al., 2022). Traditional

systems are trained for a fixed set of types, while recent models accept type names, descriptions, or annotation guidelines as part of the input (Sainz et al., 2024; Zaratiana et al., 2025). This allows users to specify what to extract without training a separate model for every schema. The field has therefore moved towards general-purpose extraction, with a single model intended to serve multiple tasks and datasets (Lu et al., 2022a; Wang et al., 2023).

Supporting a new schema does not by itself ensure accurate extraction on a new dataset. Models must transfer beyond the domains and annotation conventions represented in their training data (Bassignana and Plank, 2022; Sainz et al., 2024). Discriminative models predict spans and their types directly from the input. UniEX (Ping et al., 2023), GLiNER (Zaratiana et al., 2024), GLiNER2 (Zaratiana et al., 2025), and GLiFormer (Knowledgator Team, 2026) support flexible schemas, but their performance varies across domains and extraction tasks, as our evaluation shows. Generative models such as UIE (Lu et al., 2022b), InstructUIE (Wang et al., 2023), and GoLLiE (Sainz et al., 2024) express extraction targets through text generation. This accommodates detailed instructions, but generated outputs can introduce unsupported entities or fail to preserve the original mention boundaries. GPT-NER reports false entity predictions (Wang et al., 2025), while GNER documents boundary errors and omissions, additions, or changes to input words (Ding et al., 2024). The remaining challenge is to support user-specified schemas while maintaining extraction accuracy across domains and annotation guidelines.

To address this challenge, we investigate whether reward-based training can improve a schema-conditioned extractor beyond likelihood supervision alone. We present CALON, a unified 0.6B-parameter encoder for entity, relation, and event extraction. Drawing on RLCD’s emphasis on decisions with explicit probabilities (TypeSafe

✉ The corresponding author is Chenghua Lin, who can be contacted at chenghua.lin@manchester.ac.uk.

AI, 2026), we supplement supervised learning with optimisation of sampled extraction decisions. As in GRPO (Shao et al., 2024), multiple predictions for the same input provide a reward baseline without requiring an additional critic. For each candidate span or span pair, we sample predictions by adding Gaussian noise to the type scores. Correctly typed extractions receive positive rewards, false positives and incorrect types receive negative rewards, and predictions of no extraction receive zero reward. We compare each prediction’s reward with the average reward of the other samples and use this difference to reinforce higher-reward decisions. The resulting objective combines likelihood supervision with direct optimisation of the expected reward of sampled extractions. The motivation is to strengthen correct predictions when the model can also produce incorrect alternatives. For the same candidate, changing the type scores can lead to a correct extraction, a wrong type, or a false positive. Comparing their rewards gives an additional training signal that favours the correct extraction over these mistakes. We use this signal to supplement likelihood supervision, aiming to make the model less likely to return an incorrect prediction when its scores change.

Across 40 benchmarks, CALON outperforms the evaluated baselines on most benchmarks. It leads every reported relation extraction benchmark, including tasks represented as span-pair extraction such as aspect sentiment triplet extraction, pronoun resolution, coreference, and temporal and causal relations between events. It also achieves the highest average F1 for event triggers and arguments with gold triggers. These results demonstrate that a single extraction model can transfer across datasets and support tasks beyond conventional entity, relation, and event extraction. We make the following contributions.

1. We introduce CALON, a unified encoder for extraction under user-specified schemas. It handles entities, relations, and events through predictions over spans and span pairs. The same formulation also supports tasks beyond conventional IE (§3).
2. We develop a training objective that combines likelihood supervision with relative rewards on sampled extraction decisions. Gaussian perturbations generate predictions for each candidate, and comparison of their rewards guides the model towards correct extractions over false positives and incorrect type assignments, without an additional critic (§3.3).

2 Related Work

Extraction across schemas and data sources.

UIE (Lu et al., 2022b) unifies entity, relation, and event extraction through structured generation, while USM (Lou et al., 2023) and UniEX (Ping et al., 2023) use schema-guided matching. InstructUIE (Wang et al., 2023) uses instruction tuning, and GoLLIE (Sainz et al., 2024) incorporates annotation guidelines to support generative extraction under user-specified schemas. GLiNER (Zaratiana et al., 2024) and GLiREL (Boylan et al., 2025) match spans and span pairs, respectively, to requested types. GLiNER2 (Zaratiana et al., 2025) and GLiFormer (Knowledgator Team, 2026) support multiple tasks within one model. Training resources have also expanded through LLM annotations and synthetic examples, as in UniversalNER (Zhou et al., 2024), FiNERweb (Golde et al., 2026), and GUIDEX (Fuente et al., 2025). Combining sources requires handling differences in annotation coverage, since an unlabelled span is not necessarily a negative example (Jie et al., 2019; Li et al., 2021).

Learning extraction decisions. Span-based IE models use a null class to reject non-entity spans (Eberts and Ulges, 2020; Zhong and Chen, 2021). For generative IE, LA-RL (You et al., 2026) applies reinforcement learning through label-aware self-reflection. Inspired by RLCD (TypeSafe AI, 2026), CALON adds a reward term for sampled span and span-pair predictions. Correctly typed extractions receive positive rewards, while false positives and incorrect type assignments receive negative rewards.

Calibration. Temperature scaling can improve the calibration of neural-network predictions after training (Guo et al., 2017). Calibration can also be improved through training, with label smoothing helping pre-trained Transformers on out-of-domain data (Desai and Durrett, 2020). Even when probabilities are calibrated, selecting a threshold to maximise F1 remains dependent on the distribution of prediction scores, as shown for binary classification by Lipton et al. (2014). Class-dependent sampling can introduce a separate bias in probability estimates. Corrections for this effect have been studied in rare-event logistic regression (King and Zeng, 2001) and click prediction (He et al., 2014). CALON applies this correction principle to the lower loss weight assigned to null candidates in span and span-pair extraction.

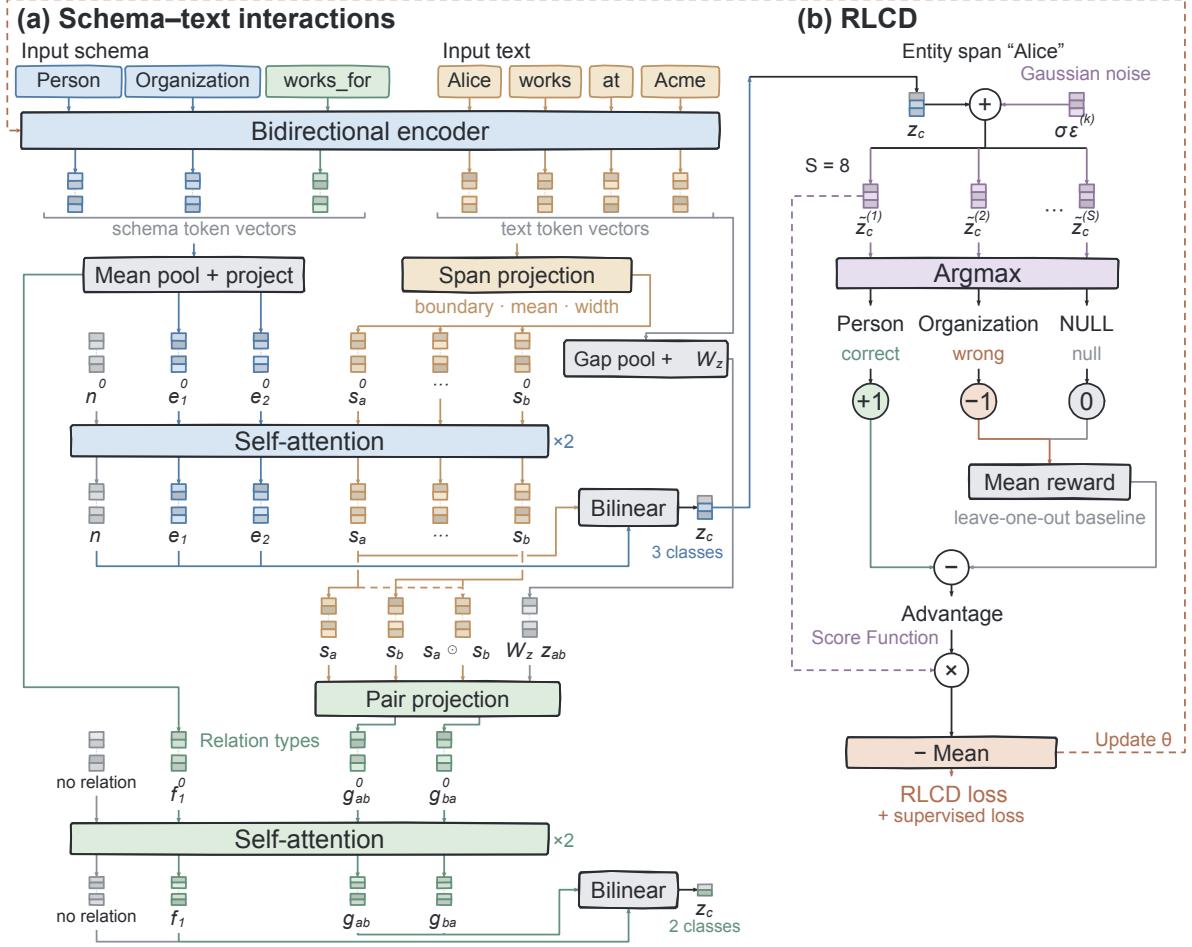


Figure 1: Overview of CALON: (a) schema-text encoding with type-candidate interaction and bilinear scoring over the queried types and a null class, and (b) the RLCD reward on sampled typed decisions added to the supervised loss.

3 Methodology

CALON jointly encodes the input text and a schema of entity and relation types with optional descriptions (§3.1). For each candidate span or ordered pair of entity spans, it predicts a probability distribution over the queried types and a null class (§3.2). The training objective combines a negative log-likelihood loss on the gold labels with a reward term based on sampled decisions (§3.3). At inference, the model selects candidates by thresholding the total probability assigned to the queried types, then labels each selected candidate with the most probable queried type (§3.4). Figure 1 illustrates this formulation.

3.1 Encoder

For entity types $\{(y_m, d_m)\}_{m=1}^M$ with descriptions d_m , relation types $\{(r_k, d'_k)\}_{k=1}^K$ and a text x , the

input is

$$\begin{aligned} &\langle \text{LABEL} \rangle y_1 : d_1 \cdots \langle \text{LABEL} \rangle y_M : d_M \\ &\langle \text{REL} \rangle r_1 : d'_1 \cdots \langle \text{REL} \rangle r_K : d'_K \langle \text{TEXT} \rangle x \end{aligned} \quad (1)$$

where $\langle \text{LABEL} \rangle$, $\langle \text{REL} \rangle$ and $\langle \text{TEXT} \rangle$ are added special tokens and a type without a description is given by its name. The encoder, Qwen3 (Qwen Team, 2025) with its causal mask removed, gives token representations $\mathbf{H} \in \mathbb{R}^{T \times D}$; the text is tokenised separately so that each token keeps its character offsets. A candidate span (i, j) of at most W text tokens is represented by its first and last token, the mean $\bar{\mathbf{h}}_{ij}$ of its tokens, and a width embedding,

$$\mathbf{s}_{ij}^0 = \mathbf{W}_s [\mathbf{H}_i; \mathbf{H}_j; \bar{\mathbf{h}}_{ij}; \mathbf{w}_{j-i}] \quad (2)$$

An entity type is represented by $\mathbf{e}_m^0 = \mathbf{W}_e \bar{\mathbf{h}}_m$, where $\bar{\mathbf{h}}_m$ is the mean over the tokens of its name and description, and a relation type by $\mathbf{f}_k^0 = \mathbf{W}_r \bar{\mathbf{h}}'_k$; all projections output vectors of size A .

Interaction layers. The score of a type for a span should depend on the other queried types and on overlapping candidates, so the types and candidates of an instance attend to each other. The sequence $[\mathbf{n}^0; \mathbf{e}_1^0; \dots; \mathbf{e}_M^0; \mathbf{s}_1^0; \dots; \mathbf{s}_C^0]$, where $\mathbf{n}^0$ is a learned vector for the null class and each element is added to a learned embedding of its kind (null, type or span), passes through L pre-norm Transformer blocks with full self-attention and no position information, which give $\mathbf{n}$, $\mathbf{e}_m$ and $\mathbf{s}_c$.

3.2 Typed Decisions with a Null Class

Label status. Training sources differ in what a missing annotation means. In an exhaustively annotated dataset such as Few-NERD (Ding et al., 2021), a queried type with no mention in a sentence is absent; in a partially annotated one, such as RAMS (Ebner et al., 2020), which labels one event per passage, it may simply not have been annotated. Each queried type in a training instance therefore has one of four statuses: *positive* (it has a gold mention), *negative* (the source states that it does not occur, e.g. types listed as absent when a synthetic text is generated), *closed-world negative* (absent because the source annotates its types exhaustively), or *unknown* (its absence cannot be inferred). Both kinds of negative are trained as absent, whereas unknown types are left out of the decision; $\mathcal{K}$ denotes the queried types whose status is not unknown.

Entities. Each candidate span c takes exactly one class from $\{0\} \cup \mathcal{K}$, where 0 is the null class, with logits

$$z_{m,c} = \frac{1}{\sqrt{r}} (\mathbf{V}\mathbf{e}_m)^\top (\mathbf{U}\mathbf{s}_c) + b \quad (3)$$

$$z_{0,c} = \frac{1}{\sqrt{r}} (\mathbf{V}\mathbf{n})^\top (\mathbf{U}\mathbf{s}_c) \quad (4)$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{r \times A}$ form a bilinear form of rank r , b is a learned bias, and the null logit depends on the span through $\mathbf{n}$. The class distribution is

$$p_c(m) = \frac{\exp z_{m,c}}{\sum_{m' \in \{0\} \cup \mathcal{K}} \exp z_{m',c}} \quad (5)$$

If an instance has an unknown type, spans without a gold type are ignored in training, since their class may be that type.

Relations. Relation candidates are the ordered pairs (a, b) , $a \neq b$, of entity spans in an instance (at most 24 spans; in training the gold entity spans, those in a gold relation first). A pair is represented

by the span vectors, their element-wise product, and the mean $\mathbf{z}_{ab}$ of the tokens between the two spans (zero if there are none),

$$\mathbf{g}_{ab}^0 = \mathbf{W}_g [\mathbf{s}_a; \mathbf{s}_b; \mathbf{s}_a \odot \mathbf{s}_b; \mathbf{W}_z \mathbf{z}_{ab}] \quad (6)$$

Each pair makes the same single choice among “no relation” and the queried relation types (Eqs. (3)-(5)), after interaction layers of its own over the no-relation vector, the relation types $\mathbf{f}_k^0$ and the pairs, with a separate bilinear form, bias and null vector. Since (a, b) and (b, a) are different candidates, the model also predicts the direction of a relation.

Events. An event mention is a trigger span of an event type with argument spans that fill its roles, and each event text gives two training instances. In the first, the event types are queried as entity types and the triggers are their spans. In the second, the event types present in the text are queried as entity types, with their triggers as spans, and their roles both as entity types, labelling the argument spans, and as relation types; each argument is a relation from its trigger to its span, labelled with its role. At inference, the entity decision first predicts triggers; a second pass over the predicted event types and their roles proposes argument spans, and the relation decision labels each (trigger, span) pair with a role of the trigger’s event type.

3.3 Training Objective

Let $\mathcal{G}_c \subseteq \mathcal{K}$ be the gold types of candidate c (a span can have a coarse and a fine type), with $\mathcal{G}_c = \{0\}$ if it has none. We maximise their total probability, $\ell_c = -\log \sum_{g \in \mathcal{G}_c} p_c(g)$ (Appendix A gives an optional Brier term). Nearly all candidates belong to the null class, so they get a weight $\omega \leq 1$, and the loss of an instance is the weighted mean

$$\mathcal{L}_{\text{inst}} = \frac{\sum_c \omega_c \ell_c}{\sum_c \omega_c} \quad (7)$$

averaged over the instances of a batch to give $\mathcal{L}_{\text{ent}}$.

Decision reward. Each candidate also acts as a stochastic policy. Its logits $\mathbf{z}_c$ over $\{0\} \cup \mathcal{K}$ are perturbed, $\tilde{\mathbf{z}}_c^{(k)} = \mathbf{z}_c + \sigma \epsilon^{(k)}$ for $k = 1, \dots, S$, and each sample takes the decision $\arg\max_m \tilde{z}_{m,c}^{(k)}$, rewarded as the evaluation scores it: $R = 1$ for a gold type, $R = -1$ for a type that is not gold, and $R = 0$ for the null class. With the advantage $\delta_c^{(k)} = R_c^{(k)} - \bar{R}_c^{(-k)}$

over the mean reward of the other samples,

$$\ell_c^{\text{PG}} = -\frac{1}{S} \sum_{k=1}^S \delta_c^{(k)} \log \pi(\tilde{\mathbf{z}}_c^{(k)} | \mathbf{z}_c) \quad (8)$$

with $\pi(\cdot | \mathbf{z}_c) = \mathcal{N}(\mathbf{z}_c, \sigma^2 \mathbf{I})$, is the score-function estimate of the negative expected reward; it is averaged with the weights of Eq. (7) and added with weight ρ . The relation loss $\mathcal{L}_{\text{rel}}$ is computed in the same way over candidate pairs, with its own null weight, and $\mathcal{L} = \mathcal{L}_{\text{ent}} + \lambda \mathcal{L}_{\text{rel}}$.

3.4 Decoding and Calibration

A span is extracted if $1 - p_c(0) \geq \tau$ and labelled with $\arg \max_{m \in \mathcal{K}} p_c(m)$; relations use the same rule, and overlapping spans are allowed. $1 - p_c(0)$ is the probability that the span belongs to any queried type, whereas the predicted type’s probability also depends on how the schema divides that mass among overlapping types (e.g., a coarse and a fine type), so thresholding it would reject spans whose evidence is shared. We therefore threshold the null class. For entities, the logits are divided by a temperature T that minimises the multi-class Brier score, and τ maximises F1, both on the calibration splits of the training sources. For relations, $T = 1$ and τ is fitted on REBEL (Huguet Cabot and Navigli, 2021) with gold entities.

Prior correction. Down-weighting null candidates by ω (Eq. (7)) shifts the minimiser of the loss. The model reports the odds of the null class against the types ω times too low, as after down-sampling negatives (King and Zeng, 2001; He et al., 2014). The correction adds $\log(1/\omega)$ to the null logit before the division by T , leaving the relative odds of the types unchanged. It needs no data and does not change the ranking of candidates at a fixed T , so after refitting T and τ it changes mainly the calibration (§5).

4 Experimental Setting

Training data. We train on 3.1M instances from 84 sources covering entities, relations, and events: human-labelled data, LLM-annotated text, and synthetic examples spanning diverse domains, genres, and schemas. We check synthetic annotations with additional models, remove near-duplicate texts, and filter training texts that substantially overlap with evaluation sentences. The training-data generation pipeline does not use evaluation-set texts or annotations as inputs, demonstrations, or style references.

No labelled examples from any target dataset are used for training, model selection, or calibration. Appendices B.1 and K describe the preparation, sampling, type distribution and domain coverage.

Evaluation. We evaluate on 40 benchmarks covering NER, RE, and EE across multiple domains. Appendix B.2 describes the datasets and scoring. We report exact-match micro F1 for each benchmark. Each benchmark receives equal weight when computing average F1. The main results use type descriptions for models that support them. Appendix D reports results with type names alone. Directed relation predictions must also match the gold direction. For RE, we report end-to-end F1 and F1 with gold entity spans. For EE, we report trigger classification F1 and argument classification F1 with gold and predicted triggers.

Baselines. We compare with GLiNER2-large (Zaratiana et al., 2025), GLiFormer-large (Knowl-edgator Team, 2026), GNER-LLaMA-7B (Ding et al., 2024) (NER only), InstructUIE (Wang et al., 2023), and GoLLIE-7B (Sainz et al., 2024), and for RE with gold entities GLiREL-large (Boylan et al., 2025), with default decoding and the same scoring code. GLiFormer, GNER, InstructUIE and GLiREL receive type names only, following their released interfaces; adding descriptions led to invalid outputs for GNER and InstructUIE. For RE with gold entities, we match the other baselines’ end-to-end relation endpoints to gold entities by span or mention text and discard unmatched relations. GLiREL receives gold entities directly (Appendix B.3).

Implementation. We use Qwen3-0.6B (Qwen Team, 2025) as our backbone to limit the computational cost of training and inference. We remove its causal mask so that text and schema tokens can attend to the full input (Li et al., 2023). We select the checkpoint, hyper-parameters, temperature, and extraction thresholds using only validation data from the training sources. Appendix B.4 provides further details.

5 Results and Discussion

Tables 1, 2, and 3 report results for entity, relation, and event extraction. We first discuss performance across tasks and domains, then examine the model components, calibration, and efficiency. Appendix C provides further discussion of individual datasets.

Dataset	CALON	GLiNER2	GLiFormer [†]	GNER [†]	InstructUIE [†]	GoLLIE
<i>CrossNER</i>						
AI	66.1	56.1	<u>59.1</u>	56.3	48.1	49.2
Literature	63.5	62.5	<u>62.8</u>	63.5	44.2	55.7
Music	<u>72.0</u>	67.7	76.2	71.1	58.8	62.8
Politics	73.0	65.2	<u>70.6</u>	63.8	45.7	47.5
Science	<u>64.4</u>	60.5	69.0	60.8	45.9	48.5
Average	67.8	62.4	<u>67.6</u>	63.1	48.5	52.7
<i>MIT</i>						
Movie	58.1	50.6	<u>57.5</u>	54.5	52.8 [‡]	52.3
Restaurant	<u>66.5</u>	55.3	61.7	46.6	80.9[‡]	24.2
Average	<u>62.3</u>	53.0	59.6	50.6	66.9[‡]	38.2
<i>Other domains</i>						
SciER	70.9	51.6	51.9	46.1	51.4	<u>54.6</u>
SciER (OOD)	67.2	<u>49.5</u>	48.4	41.7	45.8	32.8
GENTLE	42.0	24.1	<u>27.2</u>	11.3	15.5	16.0
BioRED	72.1	57.3	<u>57.5</u>	43.4	33.7	<u>60.4</u>
MeasEval	30.6	<u>22.7</u>	22.4	3.9	7.4	17.7
Average	56.6	41.1	<u>41.5</u>	29.3	30.8	36.3
<i>Personal information</i>						
ai4privacy	68.8	55.0	<u>56.1</u>	36.6	34.8	51.6
spy legal	53.2	47.4	<u>52.8</u>	20.1	50.3	44.8
spy medical	53.5	48.2	53.5	21.7	<u>49.4</u>	46.1
Average	58.5	50.2	<u>54.2</u>	26.1	44.8	47.5

Table 1: Zero-shot NER, micro F1 in %. For each sentence, every model receives the full type set of its dataset (e.g., all 14 types of CrossNER AI, including types that do not occur in the sentence), each with a short description; models marked [†] accept only type names and receive the names alone. [‡]: the model was trained on the training split of this dataset. The best score per row is in **bold**, and the second-highest distinct score is underlined.

5.1 Main Results

Entity extraction. CALON achieves the highest average F1 on CrossNER, specialised entity extraction, and personal information extraction (Table 1). Its largest gains are on scientific, biomedical, and measurement text. GLiFormer leads on CrossNER Music and Science, while InstructUIE leads on MIT Restaurant. InstructUIE’s Restaurant score raises its MIT average above CALON’s, and both MIT datasets were included in training for the evaluated InstructUIE checkpoint (Appendix B.3).

Relation extraction. CALON leads every reported RE benchmark, both with gold entities and where end-to-end scoring is available (Table 2). The same span-pair formulation also performs well on ASTE and pronoun resolution. End-to-end results are omitted for datasets that annotate relations only for selected entity pairs. GLiREL is evaluated only with gold entities because it does not support end-to-end extraction. We also compare GLiREL under two candidate-pair settings (Appendix B.3). When only pairs with an annotated relation are supplied, its average relation typing accuracy across

the 18 RE benchmarks is 40%, compared with 79% for CALON. When every pair of gold entities is considered, their average F1 scores are 2.6 and 49.8, respectively. The latter setting tests both relation typing and rejection of unrelated pairs, highlighting an important requirement for relation extraction.

Event extraction. CALON achieves the highest average F1 for trigger classification and argument classification with gold triggers (Table 3). InstructUIE’s CASIE score raises its average end-to-end argument F1 above CALON’s. The CASIE training split was included in training for the evaluated InstructUIE checkpoint (Appendix B.3). Excluding CASIE, CALON achieves the highest average F1 for all three EE metrics, with end-to-end argument F1 of 10.9 compared with InstructUIE’s 8.4. Individual dataset exceptions, including M2E2 and SPEED++ zika, are discussed in Appendix C.

5.2 Schemas and Type Descriptions

Type descriptions are particularly helpful when relation names leave their meaning unclear. On CrossRE, descriptions increase average F1 with gold entities by about 16 points, while NER group

Dataset	Gold entities						End-to-End					
	CALON	GLiNER2	GLiFormer [†]	GLiREL [†]	InstructUIE [†]	GoLLIE	CALON	GLiNER2	GLiFormer [†]	GLiREL [†]	InstructUIE [†]	GoLLIE
<i>CrossRE</i>												
AI	33.6	7.5	7.5	3.8	8.9	6.3	24.6	6.4	7.0	–	6.5	5.2
Literature	48.2	18.2	9.5	1.7	3.9	20.5	35.0	15.3	9.1	–	3.2	15.5
Music	51.6	9.5	7.9	2.4	3.9	13.6	40.1	8.3	7.6	–	3.4	10.3
News	61.9	22.1	22.2	6.7	4.8	1.5	58.4	20.0	22.0	–	3.8	1.3
Politics	40.4	5.2	7.4	3.2	2.3	6.8	32.6	4.6	7.3	–	2.0	5.0
Science	36.8	7.3	11.3	3.8	5.0	11.2	25.5	5.8	10.4	–	3.8	9.3
Average	45.4	11.6	11.0	3.6	4.8	10.0	36.0	10.1	10.6	–	3.8	7.8
<i>ASTE</i>												
14lap	75.6	2.5	0.7	0.0	8.9	0.0	43.5	2.3	0.7	–	6.5	0.0
14res	86.3	5.0	0.8	0.0	20.9	0.0	61.2	4.8	0.8	–	17.8	0.0
15res	83.9	5.1	1.2	0.0	20.0	0.0	54.5	4.8	1.2	–	15.9	0.0
16res	86.3	7.0	1.9	0.0	18.7	0.0	60.5	6.6	1.9	–	14.9	0.0
Average	83.0	4.9	1.2	0.0	17.1	0.0	55.0	4.6	1.2	–	13.7	0.0
<i>Other tasks</i>												
SciER	43.4	10.6	24.1	1.4	17.1	1.5	27.0	8.0	18.7	–	10.1	1.4
SciER (OOD)	38.9	10.3	16.1	4.5	20.7	0.7	22.0	7.6	10.9	–	10.3	0.7
BioRED	17.7	0.8	0.0	1.1	2.3	1.2	–	–	–	–	–	–
GAP	76.3	0.0	0.0	1.4	0.0	3.9	–	–	–	–	–	–
LitBank	16.8	0.0	0.0	0.1	0.0	0.0	–	–	–	–	–	–
MAVEN-ERE	2.0	0.0	0.0	0.0	0.0	0.0	–	–	–	–	–	–
MeasEval	57.9	2.1	8.9	0.0	1.0	0.8	16.6	2.0	6.2	–	1.0	0.6
Re-TACRED	38.4	18.3	13.8	16.0	11.4	20.5 [‡]	–	–	–	–	–	–
Average	36.4	5.3	7.9	3.1	6.6	3.6 [‡]	21.9	5.8	11.9	–	7.1	0.9

Table 2: RE micro F1 (%). *Gold entities* uses supplied gold spans for CALON and GLiREL and adapted baseline predictions matched to gold entities (Appendix B.3). *End-to-end* predicts entities and relations. Dashes mark unevaluated end-to-end conversions or GLiREL’s unsupported end-to-end setting. [†] marks type names alone. [‡] marks GoLLIE, which was trained on TACRED sentences reused by Re-TACRED. Best scores are bold.

Dataset	Arg. (gold triggers)				Trigger (end-to-end)				Arg. (end-to-end)			
	CALON	GLiNER2	InstructUIE [†]	GoLLIE	CALON	GLiNER2	InstructUIE [†]	GoLLIE	CALON	GLiNER2	InstructUIE [†]	GoLLIE
CASIE	37.9	3.4	62.6[‡]	3.9	33.8	7.5	67.6[‡]	10.1	14.9	3.1	49.4[‡]	3.2
MLEE	11.8	0.0	3.7	0.2	52.0	0.2	3.7	3.2	4.5	0.0	0.6	0.2
M2E2	30.9	15.5	50.2	33.6	9.2	21.4	0.0	38.3	3.6	8.1	0.0	18.3
MAILEX	23.9	0.0	4.7	0.0	9.3	0.0	0.5	0.4	2.2	0.0	0.0	0.0
SENTIVENT	37.7	0.3	28.5	3.7	14.7	0.4	11.8	8.8	7.6	0.2	5.0	3.1
SPEED++ mpox	46.9	1.3	37.5	10.5	41.0	2.8	35.7	13.2	22.1	1.1	18.3	7.6
SPEED++ zika	56.8	11.2	45.5	18.2	42.0	15.0	49.4	25.5	25.5	10.3	26.5	14.6
Average	35.1	4.5	33.2 [‡]	10.0	28.9	6.8	24.1 [‡]	14.2	11.5	3.3	14.3[‡]	6.7

Table 3: Event extraction, micro F1 in %. *Arg. (gold triggers)*: argument and role classification given the gold triggers; *Trigger* and *Arg. (end to end)*: trigger classification, and argument classification on the predicted triggers. Models marked [†] accept only type names and receive the names alone. [‡]: the model was trained on the training split of this dataset. Best score per row in **bold**.

averages change by one to three points. Descriptions can therefore provide information that the type names alone do not convey. Every entity type evaluated also occurs by name in our training data. The NER benchmarks thus test transfer across domains and annotation guidelines (Table 5, Appendix F). We find no consistent increase in per-type F1 with how often a type is annotated during training. CALON’s largest margin over the strongest baseline is observed for the least frequently annotated third of the entity types. These findings suggest that frequent type names alone do not explain its performance.

5.3 Ablation Study

Extraction and type prediction. Independent sigmoid scores reduce average NER F1 from 61.5

to 53.4, the largest NER decrease among the tested ablations (Figure 2). Thresholding the most probable type also lowers NER F1. In the softmax model, related types can share a candidate’s probability. A threshold on an individual type may therefore reject a candidate even when its total non-null probability is high. Thresholding the total non-null probability uses the evidence across types to decide whether to extract. These results support learning a joint distribution over types and a null class while separating extraction from type selection.

Architecture and training objectives. The MLP head variant slightly improves NER but lowers RE and event-argument F1. This suggests that the full prediction head is more useful for span-pair

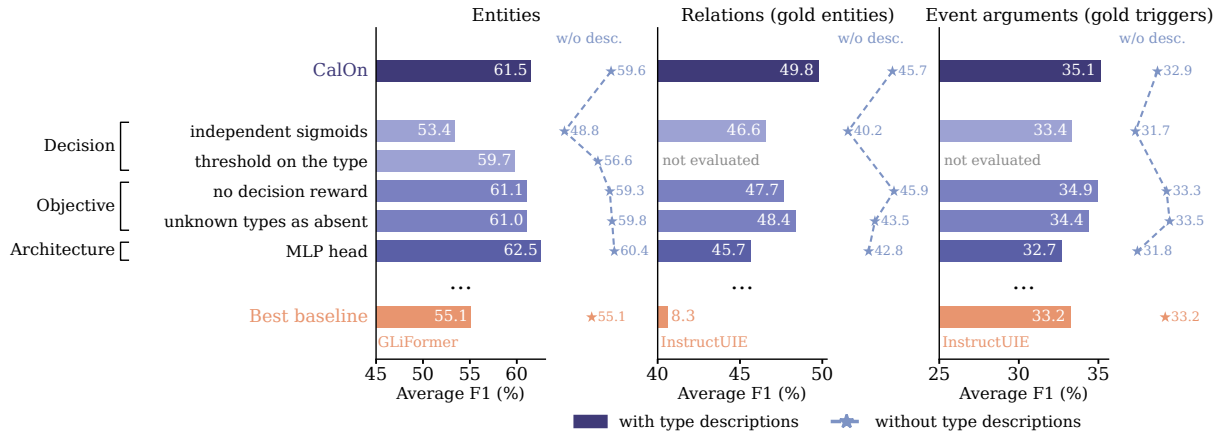


Figure 2: Ablation results for CALON. Each row changes one model component. Bars show average micro F1 (%) with type descriptions over 15 NER benchmarks, 18 RE benchmarks with gold entities, and 7 EE benchmarks with gold triggers. Stars (★) show results with type names alone on their own scales. The bottom row shows the strongest baseline for each task, averaged over the same benchmarks. InstructUIE’s training data include some of the event benchmarks.

prediction than for entity typing. With type descriptions, the decision reward mainly benefits RE, and treating unknown types as negatives lowers F1 across all three tasks. These training effects are smaller and less consistent for NER and EE without descriptions, suggesting that their benefits can depend on how the schema is presented.

5.4 Calibration

Null-class correction addresses the probability bias introduced by giving null candidates less weight in the training loss. After correction and refitting, average expected calibration error falls from 0.25 to 0.10 for relations, from 0.33 to 0.23 for entities, and from 0.51 to 0.39 for event arguments, with little change in F1 (Table 9). The results indicate more accurate confidence estimates with similar extraction accuracy. Event predictions reuse the entity and relation calibration parameters. Their calibration also improves, although event arguments retain the highest average calibration error.

5.5 Efficiency

For a given schema, CALON scores all queried types and candidates in one forward pass. Under the same PyTorch setting, it processes 120 sentences per second, 5 to 25 times the throughput of the baselines (Table 4, Appendix G). This advantage includes GLiNER2 and GLiFormer, which have similar parameter counts. CALON also achieves the highest average NER F1 among the compared models. Its peak GPU memory usage is 8.0 GB, higher than GLiNER2 and GLiFormer, so the throughput advantage is accompanied by greater runtime memory use.

Model	Params	Weights (GB)	Peak (GB)	Sent./s	NER F1
CALON	0.6B	1.2	8.0	120	61.5
GLiNER2	0.5B	1.0	2.9	22	51.6
GLiFormer	0.6B	1.2	3.2	18	55.1
GNER	6.7B	13.5	28.2	4.9	42.8
InstructUIE	11.4B	22.8	28.0	16	44.3
GoLLIE	6.7B	13.5	25.1	10.8	44.3

Table 4: Inference cost on a GH200 GPU using PyTorch and Transformers with bf16 weights. Throughput is measured on MIT Movie with all 12 types, excluding model loading. Generative models use greedy decoding with batch size 32. NER F1 is averaged over the 15 benchmarks in Table 1.

6 Conclusion

We presented CALON, a unified 0.6B-parameter encoder for extraction under user-specified schemas. The model handles entities, relations, and events through span and span-pair predictions. Its training objective combines likelihood supervision with relative rewards on sampled extractions, favouring correct predictions over false positives and incorrect type assignments. Across 40 benchmarks, CALON outperforms the evaluated baselines on most benchmarks and leads every reported relation extraction benchmark. It also achieves the highest average F1 for event triggers and arguments with gold triggers. Its advantages extend to tasks represented as span-pair extraction, including aspect sentiment triplet extraction, pronoun resolution, coreference, and temporal and causal relations between events. Probability correction and calibration improve confidence estimates across all three extraction tasks with little change in F1.

Limitations

Our training data and evaluation are limited to English, and transfer to other languages has not been tested. Candidate spans are capped at 32 tokens, and relations are predicted only between spans within the same input. These limits leave longer mentions and relations across inputs outside the model’s current scope. Much of the training data is annotated by LLMs. Voting across annotators and checks by judges help reduce shared biases but cannot fully eliminate them. Calibration parameters are fitted on validation data from the training sources, and the resulting probabilities may remain imperfect when the domain changes. We report results from a single seed because repeating training and the full evaluation across 40 benchmarks would require substantial additional computation.

Ethical Considerations

When applied to text containing personal information, extraction systems can make sensitive details easier to identify and aggregate. Such applications should respect privacy and the conditions under which the data were obtained. LLM annotations may also retain shared biases despite our annotation checks. Predictions and confidence estimates should be reviewed before they are used in decisions that affect individuals.

Acknowledgments

We used LLMs to assist with drafting and restructuring parts of the main text and appendices. The authors reviewed and revised the resulting material and take responsibility for the methods, experimental results, claims, and references in the final manuscript.

References

- Marah Abidin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Ai4Privacy. 2025. [Open pii masking 500k dataset](#).
- Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. [GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation](#). In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.
- David Bamman, Olivia Lewke, and Anya Mansoor. 2020. [An annotated dataset of coreference in English literature](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.
- Elisa Bassignana and Barbara Plank. 2022. [CrossRE: A cross-domain dataset for relation extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jack Boylan, Chris Hokamp, and Demian Gholipour Ghalandari. 2025. [GLiREL - generalist model for zero-shot relation extraction](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8230–8245, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. UniMax: Fairer and more effective language sampling for large-scale multilingual pre-training. In *The Eleventh International Conference on Learning Representations*.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.
- Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. [Few-NERD: A few-shot named entity recognition dataset](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.
- Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Yan Bowen, and Min Zhang. 2024. [Rethinking negative instances for generative named entity recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3461–3475, Bangkok, Thailand. Association for Computational Linguistics.
- Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020: 24th European Conference on Artificial Intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain-Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, pages 2006–2013. SAGE Publications 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP.

- Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. [Multi-sentence argument linking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. [MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.
- Neil De La Fuente, Oscar Sainz, Iker García-Ferrero, and Eneko Agirre. 2025. [GUIDEX: Guided synthetic data generation for zero-shot information extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24248–24262, Vienna, Austria. Association for Computational Linguistics.
- Jonas Golde, Patrick Haller, and Alan Akbik. 2026. [FiNERweb: Datasets and artifacts for scalable multilingual named entity recognition](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2281–2300, Rabat, Morocco. Association for Computational Linguistics.
- Jonas Golde, Patrick Haller, Max Ploner, Fabio Barth, Nicolaas Jedema, and Alan Akbik. 2025. [Familiarity: Better evaluation of zero-shot named entity recognition by quantifying label shifts in synthetic training data](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 820–834, Albuquerque, New Mexico. Association for Computational Linguistics.
- Thomas Green, Diana Maynard, and Chenghua Lin. 2022. [Development of a benchmark corpus to support entity recognition in job descriptions](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208, Marseille, France. European Language Resources Association.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 1321–1330. JMLR.org.
- Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. [SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.
- Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñero Candela. 2014. [Practical lessons from predicting clicks on ads at facebook](#). In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, ADKDD’14*, page 1–9, New York, NY, USA. Association for Computing Machinery.
- Dan Hendrycks and Kevin Gimpel. 2023. [Gaussian error linear units \(gelus\)](#). *Preprint*, arXiv:1606.08415.
- Pere-Lluís Hugué Cabot and Roberto Navigli. 2021. [REBEL: Relation extraction by end-to-end language generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gilles Jacobs and Véronique Hoste. 2022. Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, 56(1):225–257.
- Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. [Better modeling of incomplete annotations for named entity recognition](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.
- Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.
- Knowledgator Team. 2026. [Generalist multitask transformer encoders](#). <https://github.com/Knowledgator/GLiFormer>.
- Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. [Cross-media structured common space for multimedia event extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.
- Yangming Li, lemao liu, and Shuming Shi. 2021. [Empirical analysis of unlabeled entity problem in named entity recognition](#). In *International Conference on Learning Representations*.
- Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.
- Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 225–239. Springer.

- Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.
- Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI conference on Artificial Intelligence*.
- Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022a. [Unified structure generation for universal information extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.
- Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022b. [Unified structure generation for universal information extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.
- Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. [Biored: a rich biomedical relation extraction dataset](#). *Briefings in Bioinformatics*, 23(5):bbac282.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.
- Mistral AI. 2025. Mistral small 3. <https://mistral.ai/news/mistral-small-3>.
- NVIDIA, :, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, and 200 others. 2026. [Nemotron 3 nano omni: Efficient and open multimodal intelligence](#). *Preprint*, arXiv:2604.24954.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. [GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.
- Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. 2024. [SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, USA. Association for Computational Linguistics.
- Yang Ping, Jun Yu Lu, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Pingjian Zhang, and Jiaxing Zhang. 2023. [UniEX: An effective and efficient framework for unified information extraction via a span-extractive perspective](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16424–16440, Toronto, Canada. Association for Computational Linguistics.
- Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun’ichi Tsujii, and Sophia Ananiadou. 2012. [Event extraction across multiple levels of biological organization](#). *Bioinformatics*, 28(18):i575–i581.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. [GoLLIE: Annotation guidelines improve zero-shot information-extraction](#). In *The Twelfth International Conference on Learning Representations*.
- Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*.
- Maksim Savkin, Timur Ionov, and Vasily Kononov. 2025. [SPY: Enhancing privacy with synthetic PII detection dataset](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 236–246, Albuquerque, USA. Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang,

- Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Saurabh Srivastava, Gaurav Singh, Shou Matsumoto, Ali Raz, Paulo Costa, Joshua Poore, and Ziyu Yao. 2023. [MailEx: Email event and argument extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12964–12987, Singapore. Association for Computational Linguistics.
- George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI conference on artificial intelligence*.
- Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. [PHEE: A dataset for pharmacovigilance event extraction from text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- TypeSafe AI. 2026. TypeSafe: Reinforcement learning for calibrated decisions (RLCD). <https://typesafe.ai>. Accessed 2026-09-29.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. [GPT-NER: Named entity recognition via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiao Wang, Wei Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, J. Yang, Siyuan Li, and Chunsai Du. 2023. [Instructuie: Multi-task instruction tuning for unified information extraction](#). *ArXiv*, abs/2304.08085.
- Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. [MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. [MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4072–4091, Bangkok, Thailand. Association for Computational Linguistics.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. [Mind the GAP: A balanced corpus of gendered ambiguous pronouns](#). *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. [Position-aware tagging for aspect sentiment triplet extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.
- Xiao You, Tianwei Yan, Zixu Shan, Longyu Du, and Shan Zhao. 2026. La-rl: Label-aware self-reflection for reinforcement learning in information extraction. *arXiv preprint arXiv:2607.23420*.
- Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. [GLiNER2: Schema-driven multi-task learning for structured information extraction](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 130–140, Suzhou, China. Association for Computational Linguistics.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. [GLiNER: Generalist model for named entity recognition using bidirectional transformer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.
- Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. [SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.
- Zexuan Zhong and Danqi Chen. 2021. [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*, pages 50–61, Online. Association for Computational Linguistics.
- Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. [UniversalNER: Targeted distillation from large language models for open named entity recognition](#). In *The Twelfth International Conference on Learning Representations*.

A Method Details

Brier term. The implementation supports an optional multi-class Brier term. For a candidate c , the target vector $\mathbf{t}_c$ assigns equal probability to the gold classes in $\mathcal{G}_c$ and zero to the remaining classes. The per-candidate loss is

$$\ell_c = -\log \sum_{g \in \mathcal{G}_c} p_c(g) + \frac{\beta}{2} \|\mathbf{p}_c - \mathbf{t}_c\|_2^2. \quad (9)$$

We set $\beta = 0$ for the reported model, so this term is not included in its training loss (Appendix B.4).

B Additional Experimental Details

B.1 Data Preparation and Sampling

Sources and schemas. Human-labelled sources include Few-NERD with fine-grained entity types (Ding et al., 2021), MASSIVE with slot annotations (FitzGerald et al., 2023), and the event datasets MAVEN-Arg (Wang et al., 2024), GENEVA (Parekh et al., 2023), RAMS (Ebner et al., 2020), and PHEE (Sun et al., 2022). FiNERWeb (Golde et al., 2026) provides LLM-annotated web passages, and Pile-NER (Zhou et al., 2024) provides LLM annotations of texts sampled from the Pile.

We use Qwen3.8-27B to generate additional texts and annotations. These sources include new annotations of REBEL texts (Huguet Cabot and Navigli, 2021) and CC-News articles¹, as well as synthetic texts generated from combinations of genres, domains, and schemas. Table 13 lists the genres and domains, and Figure 5 gives the generation prompts.

Checking generated annotations. For most generated or reannotated sources, Nemotron-3-Nano-Omni-30B-A3B (NVIDIA et al., 2026) independently annotates each text. When its annotations differ from those of Qwen3.8-27B, gpt-oss-120b (OpenAI, 2025) resolves the disagreements.

For a subset of synthetic sources, four additional models annotate each text independently of the generator. These are gpt-oss-120b, Nemotron, Phi-4 (Abdin et al., 2024), and Mistral-Small-24B (Mistral AI, 2025). Together with the generator, they provide five annotations. A candidate mention, relation, trigger, or argument is accepted when a majority of the models that provide an annotation agree on it. This requires three votes when all five models participate. Remaining disagreements

are accepted or rejected only when both judges, gpt-oss-120b and Qwen3.8-27B, agree. If a disagreement remains unresolved, or no model provides an annotation for a type, we remove that type from the queried schema for the text. Retained types therefore have a known annotation status.

Duplicate removal and incomplete annotations.

Within each synthetic source, we remove near-duplicate texts using all-mpnet-base-v2 embeddings and a cosine similarity threshold of 0.9. We also compare training texts with evaluation sentences using word 5-grams. A training text is removed when their overlap covers at least 80% of the 5-grams in either text. RAMS and GENEVA annotate only some events in each passage. For these sources, event types and argument roles without an annotated occurrence are marked unknown rather than absent.

Sampling. We sample with replacement from a pool of 3.1M training instances. With 40,000 updates and a global batch size of 128, the sampling budget is 5.12M draws, including repeated samples. UniMax (Chung et al., 2023) distributes this budget as evenly as possible across sources. The allocation for each source is capped at four times its number of training instances. The resulting sampling proportions are approximately 35% entity, 41% relation, and 23% event examples.

B.2 Evaluation Datasets and Scoring

Entity extraction. The NER evaluation covers the five CrossNER domains of AI, literature, music, politics, and science (Liu et al., 2021), as well as MIT Movie and MIT Restaurant (Liu et al., 2013). It also includes SciER’s in-domain and out-of-domain test sets (Zhang et al., 2024), GENTLE (Aoyama et al., 2023), BioRED (Luo et al., 2022), and MeasEval (Harper et al., 2021). Personal information extraction is evaluated on ai4privacy (Ai4Privacy, 2025) and the legal and medical subsets of spy (Savkin et al., 2025). Table 11 lists the domains, type counts, and evaluation sizes. Each input is queried with the dataset’s full entity schema. A queried type with no gold mention is treated as absent for scoring. We split spy documents into chunks of at most 1,200 characters, with boundaries at whitespace outside entity spans.

Relation extraction. We evaluate on the six CrossRE domains (Bassignana and Plank, 2022), using the relation definitions from the paper as descriptions. Other RE datasets are SciER,

¹<https://commoncrawl.org/blog/news-dataset-available>

BioRED, MeasEval, and Re-TACRED (Stoica et al., 2021). We also represent four ASTE datasets (Xu et al., 2020) as directed relations from aspect spans to opinion spans, with sentiment as the relation type.

Additional tasks are converted to span-pair extraction. In GAP (Webster et al., 2018), a relation links the designated pronoun to a coreferent person name. In LitBank (Bamman et al., 2020), each mention is linked to the nearest preceding mention of the same entity within the input. We use all LitBank documents for evaluation. In MAVEN-ERE (Wang et al., 2022), relation candidates are event and time-expression mentions. Temporal, causal, and subevent links use representative mentions of each event within the input. Event coreference links later mentions to the first mention of the same event.

We report only the gold-entity setting for BioRED, GAP, LitBank, MAVEN-ERE, and Re-TACRED. Their converted evaluation inputs use representative concept mentions, designated mentions, or supplied event and coreference mentions. End-to-end extraction is not evaluated under these conversions. Long inputs are divided into chunks outside annotated spans, and links whose endpoints fall in different chunks are excluded for every model. The converted LitBank and MAVEN-ERE scores therefore measure span-pair extraction within an input rather than the original document-level task.

Event extraction. We evaluate on CASIE (Satyapanich et al., 2020), MLEE (Pyysalo et al., 2012), M2E2 (Li et al., 2020), MAILEX (Srivastava et al., 2023), SENTiVENT (Jacobs and Hoste, 2022), and SPEED++ (Parekh et al., 2024). For SPEED++, we use the released English files `en-mpox.json` and `en-zika.json` and retain `mpox` and `zika` as their benchmark labels. We report trigger classification and argument classification with both gold and predicted triggers.

Scoring. We compute exact-match micro F1 for each benchmark and give each benchmark equal weight in averages. An entity prediction must match both the character offsets and the type of a gold mention. A relation prediction must match the two spans and the relation type. Directed relations must also match the direction. BioRED relations are symmetric, so reversed endpoints are treated as equivalent. An event argument must match the trigger span and event type, the argument span, and the argument role.

We also measure the false-positive rate on absent types and the miss rate. The former is the fraction

of type queries with no gold occurrence for which the model returns at least one prediction. The latter is the fraction of gold mentions without an exact matching prediction.

B.3 Evaluation Adaptations

Input formats. GNER (Ding et al., 2024) was trained on Pile-NER (Zhou et al., 2024). Following its input format, we split texts into words and separate punctuation into individual tokens. For GoLLIE (Sainz et al., 2024), each type is defined as a Python class with its description in the docstring. We leave docstrings empty when evaluating without descriptions and provide no examples in either setting. GLiNER2 (Zaratiana et al., 2025) is evaluated on events through its structured extraction interface.

Relations with gold entities. CALON and GLiREL receive gold entity spans directly. For the other baselines, we match the endpoints of their end-to-end relation predictions to gold entities. We first use an exact span match. If none exists, we use the nearest gold mention with the same text, ignoring case. Relations with an unmatched endpoint are discarded. These adapted scores still depend on the baselines identifying the mentions and do not isolate relation prediction in the same way as supplying gold entities directly.

Arguments with gold triggers. For GLiNER2 and GoLLIE, we align predicted events with gold triggers of the same event type and retain the arguments of matched events. A match is accepted when the trigger spans are identical, the predicted span contains the gold span, or their trigger texts are identical. InstructUIE predicts arguments for an event type rather than for an individual trigger. We provide the gold event types and assign each predicted argument to the nearest gold trigger of that type.

GLiREL. We split texts into words and punctuation, with additional token boundaries at the start and end of every gold entity. GLiREL receives these entities and the relation type names, with hyphens and underscores replaced by spaces. It scores every ordered pair of gold entities, and we retain the highest-scoring type when its score passes the extraction threshold. Predictions assigned the *no relation* label are not returned as relations.

We select the threshold on the REBEL calibration split from $\{0.05, 0.1, 0.2, \dots, 0.9\}$. On the same split, we compare queries with and without an additional *no relation* label. We retain this label

because it gives higher F1, 18.1 compared with 17.2, and select a threshold of 0.05. The evaluated configuration therefore includes both a threshold and a label for rejecting pairs.

We separately measure relation typing when only pairs with a gold relation are supplied. Across the 18 RE benchmarks, GLiREL’s mean typing accuracy is 40.0%, compared with 79.0% for CALON when taking its most probable non-null type. With all ordered pairs of gold entities as candidates, their average F1 scores are 2.6 and 49.8. These metrics answer different questions. The first measures typing on positive pairs, while the second also requires rejecting unrelated pairs. GLiREL’s low F1 in the latter setting is associated with low precision. Its 100% typing accuracy on GAP and LitBank is expected because each has only one relation type.

Training coverage. The released training configuration associated with the evaluated InstructUIE checkpoint includes MIT Movie, MIT Restaurant, and CASIE². GoLLIE’s training includes TACRED and ACE05 (Sainz et al., 2024). Re-TACRED revises labels for TACRED sentences (Stoica et al., 2021), while M2E2 uses the ACE event types. These are different forms of prior exposure. The former involves shared source sentences, whereas the latter involves shared event types. We mark the relevant MIT, CASIE, and Re-TACRED scores with [‡]. Such exposure is relevant to interpreting the results, but the comparison does not isolate its effect from other differences between models.

B.4 Architecture and Training Settings

Architecture and candidate limits. We use bfloat16 mixed precision. Candidate spans contain at most $W = 32$ tokens. Input texts contain at most 1,024 tokens, and the combined text and schema contain at most 3,072 tokens. The projection dimension is $A = 256$, and width embeddings have dimension 32. We use $L = 2$ interaction blocks with four attention heads each, a feed-forward dimension of $4A$, GELU activation (Hendrycks and Gimpel, 2023), and dropout of 0.1. The bilinear scoring dimension is $r = 64$. Relation candidates are formed from at most 24 entity spans per instance.

Optimisation. We train with AdamW optimiser (Loshchilov and Hutter, 2019). The learning rate is 4×10^{-5} for the encoder and 2×10^{-4} for

²https://github.com/BeyondrXX/InstructUIE/blob/master/configs/multi_task_configs/train_tasks.json

the prediction heads. Weight decay is 0.01, and gradients are clipped at 1.0. The schedule uses 10% warm-up followed by linear decay. Training runs for 40,000 updates on 32 GPUs with a global batch size of 128, or four instances per GPU. Each type description is omitted independently with probability 0.3. We evaluate every 500 updates on validation data from the training sources, using up to 1,000 instances per source. We retain the checkpoint with the highest mean F1 across these sources and use validation loss to break ties.

Loss settings. Using the same validation data, we select $\beta = 0$ and a null-class weight of $\omega = 0.1$. The latter is compared with 0.03 in a shorter preliminary run. The relation null-class weight is also 0.1, and the relation loss coefficient is $\lambda = 1$. Both entity and relation decision rewards use $S = 8$ samples, noise scale $\sigma = 1$, and coefficient $\rho = 0.5$. We increase ρ linearly from zero to 0.5 during the first 200 updates.

Calibration settings. For the main results, we fit the entity temperature by minimising the multi-class Brier score on calibration data from the training sources. We search 57 log-spaced temperatures from 0.25 to 4. Extraction thresholds are selected to maximise F1 on a grid from 0.02 to 0.98. The reported uncorrected model uses entity temperature $T = 1.219$ and threshold $\tau = 0.88$. Relations use $T = 1$ and threshold $\tau = 0.85$, fitted on REBEL with gold entities. Event triggers reuse the entity parameters, and event arguments reuse the relation parameters. Appendix E reports the additional correction and refitting analysis.

C Additional Results

Entities. CALON has the highest average F1 on CrossNER, the other-domain group, and personal information extraction (Table 1). GLiFormer remains stronger on CrossNER Music and Science. InstructUIE leads on MIT Restaurant with 80.9 F1, compared with 66.5 for CALON. This result raises its MIT average above CALON’s, but its advantage does not extend to the other NER groups. Both MIT datasets appear in the released training configuration associated with the evaluated InstructUIE checkpoint (Appendix B.3).

Relations. CALON leads every reported RE benchmark in both the gold-entity and available end-to-end settings (Table 2). This includes conventional relations, ASTE, and pronoun resolution. Its end-to-end CrossRE average is 36.0 F1, compared

with 10.6 for the strongest baseline. Scores with gold entities are zero for several baselines on GAP, LitBank, and MAVEN-ERE. However, CALON’s scores on the latter two also remain low, so its lead does not imply that these tasks are solved. Dashes indicate an unevaluated end-to-end conversion or GLiREL’s unsupported end-to-end setting, as described in Appendix B.2. The GLiNER2 paper reports NER and text-classification benchmarks but no RE or EE benchmark results³.

Events. CALON has the highest average trigger F1 and argument F1 with gold triggers (Table 3). It leads five of the seven datasets on arguments with gold triggers. InstructUIE’s CASIE end-to-end argument score of 49.4, compared with 14.9 for CALON, raises its overall average above CALON’s. CASIE appears in its released training configuration. Excluding CASIE, CALON has the highest average for all three EE metrics. Its gold-trigger argument F1 is 34.7 compared with InstructUIE’s 28.4, and its end-to-end argument F1 is 10.9 compared with 8.4.

Individual exceptions remain. InstructUIE is stronger on M2E2 arguments with gold triggers and on SPEED++ zika triggers and end-to-end arguments. GoLLIE leads on M2E2 triggers and end-to-end arguments. InstructUIE scores zero on M2E2 triggers and end-to-end arguments and on MAILEX end-to-end arguments. GLiNER2 scores zero on MLEE and MAILEX arguments with gold triggers. These differences show why dataset-level results and training coverage should accompany task averages.

D Results without Type Descriptions

Tables 6, 7, and 8 repeat the evaluation with type names alone. Models that support descriptions therefore receive the same schema information as models that accept only names. CALON is trained with each description omitted independently with probability 0.3 (Appendix B.4).

For NER, CALON retains the highest averages on the other-domain group and personal information extraction. These are 53.7 and 57.2 F1, compared with the strongest baseline averages of 41.5 and 54.2. GLiFormer leads on CrossNER with 67.6 F1, compared with 66.2 for CALON. For RE, CALON leads every reported benchmark in both settings. Its

³A GLiNER2 maintainer reports that fewer than 5% of its data contain relation annotations at <https://github.com/fastino-ai/GLiNER2/issues/3>. The full training data for the evaluated checkpoint are unavailable to us. We therefore evaluate the released weights rather than reproduce its training run.

Model	Training mentions of the name			ρ
	Rare	Middle	Frequent	
CALON	58.8	54.8	61.3	0.04
GLiNER2	50.6	46.7	52.0	0.01
GLiFormer	50.0	50.6	55.2	0.09
GNER	36.1	42.7	48.7	0.21
InstructUIE	38.7	39.5	44.2	0.10
GoLLIE	37.9	39.5	48.4	0.16

Table 5: Macro F1 (%) over the 150 entity types in the 15 NER benchmarks, counting types separately for each benchmark. Types are grouped into thirds by their mention counts in CALON’s training data. Rare types have at most 13,400 mentions, and frequent types have more than 77,924. ρ is the Spearman correlation between these counts and per-type F1. Evaluation settings follow Table 1.

gold-entity CrossRE average falls from 45.4 with descriptions to 29.1 without them, indicating that the relation definitions provide useful information beyond the names.

For EE, CALON retains the highest average trigger F1. Its argument F1 with gold triggers falls from 35.1 to 32.9, slightly below InstructUIE’s 33.2. InstructUIE also retains the highest end-to-end argument average. These results show that the effect of descriptions varies across tasks and datasets.

E Calibration per Benchmark

Table 9 reports F1 and expected calibration error (ECE) for the original fitted model, null-class correction with threshold refitting, and correction with temperature and threshold refitting. Entity calibration uses validation data from the training sources. Relation calibration uses the REBEL calibration split. The default relation decoder uses $T = 1$, while the temperature-refitting condition is an additional calibration analysis. Event triggers reuse the entity parameters, and event arguments reuse the relation parameters. No event-specific parameters are fitted.

ECE is computed over returned predictions using 15 equal-width confidence bins. The confidence is the probability assigned to the predicted type or argument role. Correctness requires the exact span and type or role matches described in Appendix B.2. RE is evaluated with gold entities, and event arguments are evaluated with gold triggers. This ECE measures the confidence of returned predictions rather than calibration over all null and non-null candidates. Correction and temperature refitting reduce average ECE from 0.33 to 0.23 for entities, from 0.25 to 0.09 for relations, and from

Dataset	CALON	GLiNER2	GLiFormer	GNER	InstructUIE	GoLLIE
<i>CrossNER</i>						
AI	63.3	51.0	<u>59.1</u>	56.3	48.1	37.5
Literature	<u>63.2</u>	57.9	<u>62.8</u>	63.5	44.2	51.0
Music	<u>71.4</u>	64.0	76.2	71.1	58.8	56.3
Politics	72.5	66.8	<u>70.6</u>	63.8	45.7	38.2
Science	<u>60.8</u>	53.6	69.0	<u>60.8</u>	45.9	42.0
Average	<u>66.2</u>	58.7	67.6	63.1	48.5	45.0
<i>MIT</i>						
Movie	<u>56.8</u>	45.1	57.5	54.5	52.8 [‡]	49.5
Restaurant	<u>65.8</u>	45.0	61.7	46.6	80.9[‡]	16.8
Average	<u>61.3</u>	45.0	59.6	50.6	66.9[‡]	33.1
<i>Other domains</i>						
SciER	67.8	37.6	<u>51.9</u>	46.1	51.4	17.6
SciER (OOD)	60.9	37.9	<u>48.4</u>	41.7	45.8	14.4
GENTLE	38.1	27.1	<u>27.2</u>	11.3	15.5	5.4
BioRED	70.9	47.8	<u>57.5</u>	43.4	33.7	57.2
MeasEval	31.1	13.8	<u>22.4</u>	3.9	7.4	1.5
Average	53.7	32.9	<u>41.5</u>	29.3	30.8	19.2
<i>Personal information</i>						
ai4privacy	66.0	<u>56.9</u>	56.1	36.6	34.8	46.8
spy legal	<u>52.6</u>	48.2	52.8	20.1	50.3	42.4
spy medical	<u>52.8</u>	48.7	53.5	21.7	49.4	44.8
Average	57.2	51.3	<u>54.2</u>	26.1	44.8	44.7

Table 6: NER with type names alone, reported as micro F1 (%). Evaluation settings follow Table 1. [‡] marks scores for the evaluated InstructUIE checkpoint whose released training configuration includes the corresponding dataset. Best scores are **bold**; second-highest distinct scores are underlined.

0.51 to 0.39 for event arguments, with little change in F1. Relation ECE decreases on 17 of the 18 benchmarks. CrossRE News is the exception, with an initial ECE of 0.08. On the REBEL calibration split, ECE is 0.13 before correction, 0.01 after correction and threshold refitting, and 0.03 after additional temperature refitting. Thus, temperature refitting does not improve every individual result.

Argument ECE decreases on all seven event benchmarks. MLEE improves from 0.29 to 0.11, but ECE remains at least 0.30 on the other six datasets. The multi-class Brier score over all relation candidates changes little, with averages of 0.157, 0.162, and 0.158 across the three settings. Since most candidate pairs are null, this metric emphasises a different set of decisions from ECE over returned relations.

F Type Familiarity

Golde et al. (2025) study zero-shot NER in terms of type-name similarity and training frequency. All entity type names in our 15 NER benchmarks occur verbatim in CALON’s training data. These benchmarks therefore test transfer across domains and annotation guidelines rather than to unseen entity

type names.

We examine whether performance depends on how frequently a type name is annotated in training. The 150 evaluated entity types, counted separately for each benchmark, are divided into thirds using their mention counts in CALON’s training data (Table 5). We report macro F1 over types in each group and the Spearman correlation between training frequency and per-type F1.

CALON has the highest macro F1 in all three groups. Its correlation with training frequency is small, $\rho = 0.04$, and its largest margin is in the rarest group, with 58.8 F1 compared with GLiNER2’s 50.6. The generative baselines have positive correlations from 0.10 to 0.21 under the same grouping. However, these frequencies come from CALON’s training data and do not measure the baselines’ own training exposure. The analysis shows that CALON’s gains are not confined to its most frequent type names. It does not isolate the effects of the training data and model architecture.

G Inference Cost

Table 4 compares inference on one GH200 GPU, with one model loaded at a time. All models use

Dataset	Gold entities						End-to-end					
	CALON	GLiNER2	GLiFormer	GLiREL	InstructUIE	GoLLIE	CALON	GLiNER2	GLiFormer	GLiREL	InstructUIE	GoLLIE
<i>CrossRE</i>												
AI	28.4	6.5	7.5	3.8	8.9	0.0	20.0	4.3	7.0	–	6.5	0.0
Literature	23.0	5.0	9.5	1.7	3.9	0.6	17.6	3.5	9.1	–	3.2	0.6
Music	24.0	2.2	7.9	2.4	3.9	0.0	18.9	1.7	7.6	–	3.4	0.0
News	41.4	10.4	22.2	6.7	4.8	0.0	36.7	8.1	22.0	–	3.8	0.0
Politics	29.5	2.8	7.4	3.2	2.3	0.8	24.9	2.1	7.3	–	2.0	0.8
Science	28.6	7.0	11.3	3.8	5.0	0.3	20.3	4.9	10.4	–	3.8	0.3
Average	29.1	5.6	11.0	3.6	4.8	0.3	23.0	4.1	10.6	–	3.8	0.3
<i>ASTE</i>												
14lap	70.2	23.6	0.7	0.0	8.9	0.0	24.9	20.8	0.7	–	6.5	0.0
14res	83.0	38.4	0.8	0.0	20.9	0.0	45.0	34.6	0.8	–	17.8	0.0
15res	81.2	44.0	1.2	0.0	20.0	0.0	40.6	37.5	1.2	–	15.9	0.0
16res	85.9	41.2	1.9	0.0	18.7	0.0	42.3	35.6	1.9	–	14.9	0.0
Average	80.1	36.8	1.2	0.0	17.1	0.0	38.2	32.1	1.2	–	13.7	0.0
<i>Other tasks</i>												
SciER	46.7	14.0	24.1	1.4	17.1	0.6	28.7	8.6	18.7	–	10.1	0.6
SciER (OOD)	45.6	13.4	16.1	4.5	20.7	0.7	24.6	6.1	10.9	–	10.3	0.7
BioRED	22.8	1.2	0.0	1.1	2.3	0.5	–	–	–	–	–	–
GAP	84.6	0.0	0.0	1.4	0.0	0.0	–	–	–	–	–	–
LitBank	24.3	0.0	0.0	0.1	0.0	0.0	–	–	–	–	–	–
MAVEN-ERE	7.0	0.0	0.0	0.0	0.0	0.0	–	–	–	–	–	–
MeasEval	58.5	2.5	8.9	0.0	1.0	0.8	15.3	1.7	6.2	–	1.0	0.6
Re-TACRED	37.8	20.1	13.8	16.0	11.4	16.5 [‡]	–	–	–	–	–	–
Average	40.9	6.4	7.9	3.1	6.6	2.4 [‡]	22.8	5.5	11.9	–	7.1	0.6

Table 7: RE with type names alone, reported as micro F1 (%). Evaluation settings follow Table 2. Dashes mark unevaluated end-to-end conversions or GLiREL’s unsupported end-to-end setting. [‡] marks GoLLIE, which was trained on TACRED sentences reused by Re-TACRED. Best scores are **bold**.

Dataset	Arg. (gold triggers)				Trigger (end to end)				Arg. (end to end)			
	CALON	GLiNER2	InstructUIE	GoLLIE	CALON	GLiNER2	InstructUIE	GoLLIE	CALON	GLiNER2	InstructUIE	GoLLIE
CASIE	33.1	0.3	62.6[‡]	2.1	31.8	0.6	67.6[‡]	4.1	12.0	0.1	49.4[‡]	1.4
MLEE	12.1	0.1	3.7	0.2	47.1	0.0	3.7	1.9	4.1	0.0	0.6	0.1
M2E2	34.2	3.3	50.2	27.4	14.3	1.3	0.0	30.9	7.3	0.5	0.0	15.6
MAILEX	19.8	0.0	4.7	0.0	10.5	0.0	0.5	0.0	2.0	0.0	0.0	0.0
SENTIVENT	31.4	0.9	28.5	5.2	17.6	1.9	11.8	7.5	7.3	0.6	5.0	4.3
SPEED++ mpox	45.7	1.1	37.5	7.0	40.7	0.4	35.7	8.0	22.7	0.3	18.3	4.3
SPEED++ zika	53.7	0.0	45.5	17.9	41.1	0.0	49.4	20.8	25.5	0.0	26.5	12.3
Average	32.9	0.8	33.2[‡]	8.5	29.0	0.6	24.1 [‡]	10.5	11.6	0.2	14.3[‡]	5.4

Table 8: EE with event and role names alone, reported as micro F1 (%). Evaluation settings follow Table 3. [‡] marks scores for the evaluated InstructUIE checkpoint whose released training configuration includes CASIE. Best scores are **bold**.

PyTorch and Transformers with bf16 weights. We time the 1,953 MIT Movie test sentences with all 12 entity types and exclude model loading. Generative models use greedy decoding and batches of 32. Their MIT Movie F1 is unchanged from the main evaluation.

CALON processes 120 sentences per second. This is about five to eight times the throughput of GLiNER2, GLiFormer, and InstructUIE, 11 times that of GoLLIE, and 25 times that of GNER. Its weight memory is 1.2 GB, compared with 13.5 GB for the 7B models. Peak GPU memory is 8.0 GB, higher than GLiNER2 and GLiFormer but lower than the generative baselines. Parameter count alone therefore does not capture the memory cost of inference. These timings depend on the inference framework. With vLLM, GNER

and GoLLIE reach 112 and 238 sentences per second, respectively. This setting uses continuous batching and shared-prefix caching and reserves most of the GPU memory. GoLLIE is faster than CALON under this separate setup, so the PyTorch comparison does not establish a speed advantage across serving frameworks. Our CALON evaluation also includes calibration-statistic computation and is not optimised solely for throughput. The table reports the cost of this evaluation implementation.

H Data Scaling

We train on nested random subsets of 100K, 250K, 500K, 1M, and 2M instances and compare them with the full 3.1M-instance model. Each subset covers all 84 sources and preserves their proportions in the full sampling mixture. Smaller subsets are

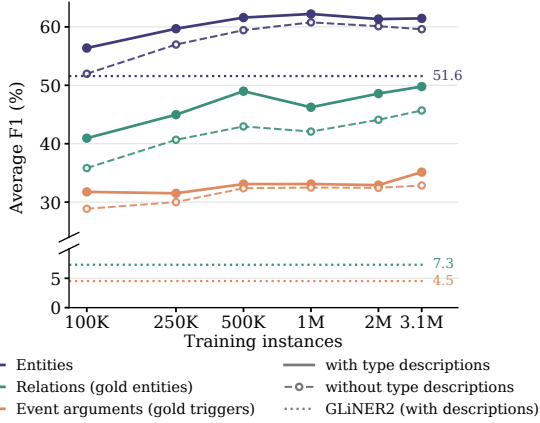


Figure 3: Average micro F1 as training data increase. Solid lines use type descriptions, dashed lines use type names alone, and dotted lines show GLiNER2 with descriptions. Averages cover the 15 NER, 18 RE with gold entities, and 7 EE with gold triggers benchmarks.

contained in larger ones. Architecture, objective, batch size, and checkpoint selection follow the full model. Training updates scale with dataset size, from 1,282 for 100K instances to 40,000 for the full data, so each run makes approximately the same number of passes over its data. Evaluation intervals are also scaled to give 80 evaluations per run. Each model is trained with one seed.

Figure 3 reports average NER F1, RE F1 with gold entities, and event-argument F1 with gold triggers. With descriptions, entity and relation averages improve by approximately five and eight points from 100K to 500K instances. Gains after 500K are smaller and not monotonic. Event-argument F1 changes less across the smaller subsets and improves by about two points with the full data. Results with type names alone show a similar pattern. Since training updates increase with dataset size, the experiment varies data and training compute together. The single-seed results do not establish whether the smaller differences are statistically reliable.

I Models

Table 10 identifies the checkpoints used for the CALON backbone, baselines, data generation and annotation, and text-embedding analysis. It also lists their model links and reported licenses.

J Evaluation Datasets

Table 11 lists the 40 evaluation benchmarks with their tasks, domains, genres, type counts, and evaluation sizes. Example counts refer to the inputs after conversion to our evaluation format,

which may be sentences, passages, or document chunks. The task conversions and scoring rules are described in Appendix B.2.

K Training Data

Table 12 lists training sources that use real texts and identifies whether their annotations come from humans, released LLM annotations, or our annotation pipeline. Counts refer to the training instances used in this work rather than the sizes of the original datasets. Table 13 lists the genres and domains used for synthetic generation. Each request combines one genre with one domain. The prompts in Figure 5 ask the generator to write a text, define an extraction schema, annotate the text, and list additional types that do not occur. These absent types provide explicit negative supervision after annotation checking.

Type distribution. Figure 4 shows type-name frequencies before training-source sampling. Human-labelled sources have fixed inventories, with 74 entity types in Few-NERD, 55 slots in MASSIVE, and between two and 162 event types in the event datasets. Pile-NER and the re-annotated REBEL texts have broader inventories with many infrequent type names. The pooled synthetic sources contain approximately 105K entity type names, 54K relation type names, and 85K event type names. These counts describe names used in the annotations and do not assume that every distinct name denotes a distinct semantic category.

Domains and genres. We sample 3,000 texts per training source and up to 500 per evaluation set. Texts are encoded with the all-mpnet-base-v2 checkpoint using the SentenceTransformers framework (Reimers and Gurevych, 2019) and projected into two dimensions with UMAP. Synthetic texts occupy a broader range of regions than the real-text training sources in this projection. Several evaluation regions also contain synthetic texts, including regions associated with reviews, search queries, social media, and personal information. Figure 7 shows how these regions relate to the assigned domains and genres. This is a qualitative view of the sampled embeddings. Overlap in the projection does not establish equivalent data distributions or explain the performance gains on its own.

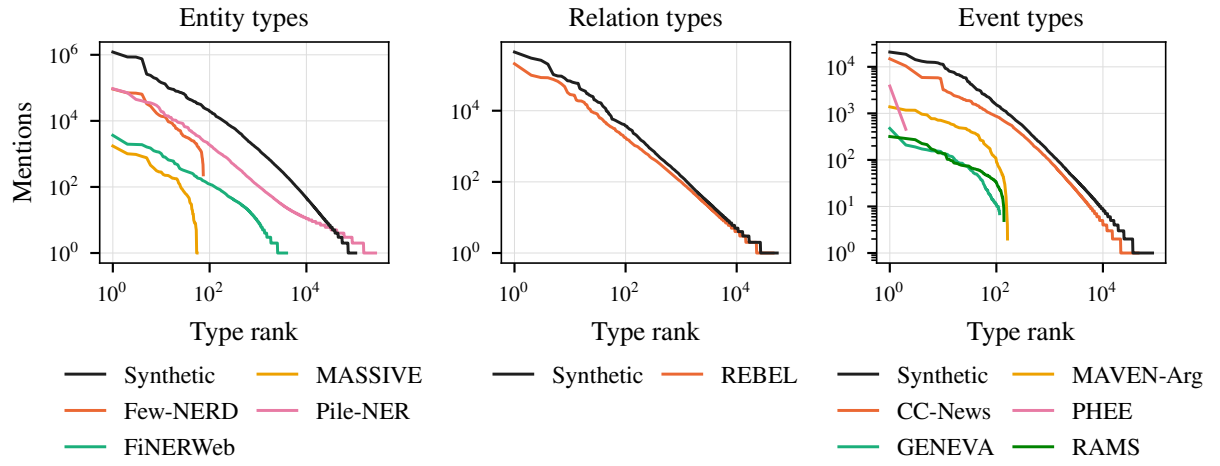


Figure 4: Training type frequencies. For each source, types are ranked by mention count. Both axes use logarithmic scales. Longer curves indicate more type names, and flatter curves indicate a more even distribution of mentions. The synthetic curve pools all synthetic sources, while the other curves show individual datasets. Counts use the complete training files before sampling.

Dataset	F1 (%)			ECE		
	Fitted	+Corr.	+Corr.+ T	Fitted	+Corr.	+Corr.+ T
<i>Entities</i>						
Validation (training sources)	68.4	68.5	68.4	0.25	0.16	0.12
CrossNER AI	66.1	66.0	65.8	0.26	0.21	0.17
CrossNER Literature	63.5	63.3	63.1	0.27	0.22	0.18
CrossNER Music	72.0	71.7	71.4	0.18	0.15	0.10
CrossNER Politics	73.0	72.8	72.6	0.23	0.18	0.15
CrossNER Science	64.4	64.4	64.3	0.21	0.17	0.12
MIT Movie	58.1	58.1	57.6	0.40	0.32	0.28
MIT Restaurant	66.5	66.5	66.2	0.31	0.23	0.20
SciER	70.9	70.7	70.7	0.24	0.15	0.12
SciER (OOD)	67.2	67.1	67.1	0.24	0.14	0.11
GENTLE	42.0	42.1	42.2	0.41	0.32	0.28
BioRED	72.1	72.0	71.9	0.21	0.13	0.10
MeasEval	30.6	30.3	29.8	0.62	0.49	0.45
ai4privacy	68.8	69.0	68.8	0.20	0.16	0.12
spy legal	53.2	53.1	53.0	0.59	0.56	0.54
spy medical	53.5	53.3	53.3	0.59	0.55	0.53
Average	61.5	61.4	61.2	0.33	0.27	0.23
<i>Relations (gold entities)</i>						
Validation (training sources)	82.8	82.8	82.9	0.13	0.01	0.03
CrossRE AI	33.6	33.7	33.5	0.49	0.27	0.24
CrossRE Literature	48.2	48.4	48.2	0.35	0.12	0.09
CrossRE Music	51.6	51.8	51.4	0.26	0.08	0.06
CrossRE News	61.9	62.2	62.8	0.08	0.10	0.13
CrossRE Politics	40.4	40.4	40.2	0.28	0.13	0.11
CrossRE Science	36.8	36.9	36.6	0.33	0.15	0.11
SciER	43.4	43.8	43.1	0.19	0.13	0.12
SciER (OOD)	38.9	38.7	38.9	0.23	0.07	0.09
ASTE 14lap	75.6	75.6	75.4	0.13	0.09	0.09
ASTE 14res	86.3	86.3	86.3	0.07	0.04	0.04
ASTE 15res	83.9	83.8	83.8	0.11	0.09	0.07
ASTE 16res	86.3	86.2	86.4	0.10	0.08	0.07
BioRED	17.7	18.0	17.8	0.30	0.10	0.09
GAP	76.3	76.4	76.2	0.11	0.05	0.07
LitBank	16.8	17.3	16.1	0.40	0.04	0.03
MAVEN-ERE	2.0	2.2	2.1	0.45	0.15	0.13
MeasEval	57.9	58.2	57.9	0.29	0.06	0.05
Re-TACRED	38.4	38.7	38.2	0.26	0.16	0.12
Average	49.8	49.9	49.7	0.25	0.11	0.09
<i>Event arguments (gold triggers)</i>						
CASIE	37.9	38.1	38.2	0.59	0.48	0.45
MLEE	11.8	12.0	11.9	0.29	0.11	0.11
M2E2	30.9	30.7	30.6	0.72	0.67	0.66
MAILEX	23.9	24.5	24.4	0.58	0.47	0.44
SENTIVENT	37.7	38.0	37.8	0.56	0.46	0.44
SPEED++ mpox	46.9	46.9	46.8	0.48	0.35	0.33
SPEED++ zika	56.8	57.0	56.6	0.37	0.31	0.30
Average	35.1	35.3	35.2	0.51	0.41	0.39

Table 9: F1 and ECE for CALON. *Fitted* uses the original calibration parameters without null-class correction. *+Corr.* applies the correction and refits thresholds with $T = 1$. *+Corr.+ T* additionally refits temperature. Parameters are selected on validation data from the training sources, using REBEL for relations and event arguments, and remain fixed on evaluation datasets. Relations use gold entities, and event arguments use gold triggers. ECE uses 15 equal-width bins over returned predictions and the probability of the predicted type or role. Lowest ECE values are **bold**.

Model	Architecture	Model link	License	Release
Qwen3-0.6B	Dense	Qwen/Qwen3-0.6B	Apache 2.0	Apr 2025
GLiNER2-large	Dense	fastino/gliner2-large-v1	Apache 2.0	Jul 2025
GLiFormer-large	Dense	knowledgator/gliformer-large-v1	Apache 2.0	Sep 2026
GLiREL-large	Dense	jackboyla/gliREL-large-v0	CC BY-NC-SA 4.0	Jan 2025
GNER-LLaMA-7B	Dense	dyyyyyyy/GNER-LLaMA-7B	Apache 2.0	Feb 2024
InstructUIE	Dense	ZWK/InstructUIE	OpenRAIL	May 2023
GoLLIE-7B	Dense	HiTZ/GoLLIE-7B	Llama 2 Community	Sep 2023
Qwen3.8-27B	Dense	Qwen/Qwen3.8-27B	Apache 2.0	Aug 2026
gpt-oss-120b	MoE	openai/gpt-oss-120b	Apache 2.0	Aug 2025
Nemotron-3-Nano-Omni-30B-A3B	MoE	nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16	NVIDIA Open Model	Apr 2026
Phi-4	Dense	microsoft/phi-4	MIT	Dec 2024
Mistral-Small-24B	Dense	mistralai/Mistral-Small-24B-Instruct-2501	Apache 2.0	Jan 2025
all-mpnet-base-v2	Dense	sentence-transformers/all-mpnet-base-v2	Apache 2.0	Mar 2022

Table 10: Model checkpoints used in this work, with their architecture, Hugging Face links, reported licenses, and release months.

Dataset	Task	Domain and genre	# Types	# Examples
CrossNER AI (Liu et al., 2021)	Entities	AI, encyclopedia	14	431
CrossNER Literature (Liu et al., 2021)	Entities	Literature, encyclopedia	12	416
CrossNER Music (Liu et al., 2021)	Entities	Music, encyclopedia	13	465
CrossNER Politics (Liu et al., 2021)	Entities	Politics, encyclopedia	9	651
CrossNER Science (Liu et al., 2021)	Entities	Science, encyclopedia	17	543
MIT Movie (Liu et al., 2013)	Entities	Film, search queries	12	1,953
MIT Restaurant (Liu et al., 2013)	Entities	Food, search queries	8	1,521
SciER (Zhang et al., 2024)	Entities	AI, scientific papers	3	854
SciER (OOD) (Zhang et al., 2024)	Entities	Science, scientific papers	3	580
GENTLE (Aoyama et al., 2023)	Entities	Mixed, unusual genres	10	1,334
BioRED (Luo et al., 2022)	Entities	Biomedicine, abstracts	6	118
MeasEval (Harper et al., 2021)	Entities	Science, scientific papers	4	149
ai4privacy (Ai4Privacy, 2025)	Entities	Personal information, forms and messages	25	3,000
spy legal (Savkin et al., 2025)	Entities	Personal information, legal documents	7	12,208
spy medical (Savkin et al., 2025)	Entities	Personal information, medical documents	7	11,785
CrossRE AI (Bassignana and Plank, 2022)	Relations	AI, encyclopedia	17	431
CrossRE Literature (Bassignana and Plank, 2022)	Relations	Literature, encyclopedia	17	416
CrossRE Music (Bassignana and Plank, 2022)	Relations	Music, encyclopedia	17	399
CrossRE News (Bassignana and Plank, 2022)	Relations	General, news	17	400
CrossRE Politics (Bassignana and Plank, 2022)	Relations	Politics, encyclopedia	17	400
CrossRE Science (Bassignana and Plank, 2022)	Relations	Science, encyclopedia	17	400
SciER (Zhang et al., 2024)	Relations	AI, scientific papers	9	854
SciER (OOD) (Zhang et al., 2024)	Relations	Science, scientific papers	9	580
ASTE 14lap (Xu et al., 2020)	Relations	Laptops, reviews	3	328
ASTE 14res (Xu et al., 2020)	Relations	Restaurants, reviews	3	492
ASTE 15res (Xu et al., 2020)	Relations	Restaurants, reviews	3	322
ASTE 16res (Xu et al., 2020)	Relations	Restaurants, reviews	3	326
BioRED (Luo et al., 2022)	Relations	Biomedicine, abstracts	8	140
GAP (Webster et al., 2018)	Relations (pronouns)	People, encyclopedia	1	2,000
LitBank (Bamman et al., 2020)	Relations (coreference)	Fiction, novels	1	1,388
MAVEN-ERE (Wang et al., 2022)	Relations (events)	History, encyclopedia	10	1,250
MeasEval (Harper et al., 2021)	Relations	Science, scientific papers	3	149
Re-TACRED (Stoica et al., 2021)	Relations	People and organisations, news and web	38	13,418
CASIE (Satyapanich et al., 2020)	Events	Cybersecurity, news	5	1,492
MLEE (Pyysalo et al., 2012)	Events	Biomedicine, abstracts	22	98
M2E2 (Li et al., 2020)	Events	Politics and conflict, news	8	5,996
MAILEX (Srivastava et al., 2023)	Events	Business, emails	10	406
SENTiVENT (Jacobs and Hoste, 2022)	Events	Finance, news	18	727
SPEED++ mpox (Parekh et al., 2024)	Events	Epidemics, social media	7	286
SPEED++ zika (Parekh et al., 2024)	Events	Epidemics, social media	7	299

Table 11: The 40 evaluation benchmarks with their tasks, domains, genres, type counts, and numbers of converted evaluation inputs. Types are entity, relation, or event types according to the task. Inputs may be sentences, passages, or document chunks. All scores use micro F1.

Dataset	Task	Domain and genre	Annotation	# Types	# Examples
Few-NERD (Ding et al., 2021)	Entities	General, encyclopedia	Human	74	125,839
MASSIVE (FitzGerald et al., 2023)	Entities (slots)	Personal assistant, voice commands	Human	55	11,062
FiNERWeb (Golde et al., 2026)	Entities	General, web pages	LLM (released)	3,895	16,798
Pile-NER (Zhou et al., 2024)	Entities	General, mixed text	LLM (released)	268,736	73,395
REBEL (Huguet Cabot and Navigli, 2021)	Entities and relations	General, encyclopedia	LLM (ours)	45,074	399,983
MAVEN-Arg (Wang et al., 2024)	Events	General, encyclopedia	Human	162	12,889
GENEVA (Parekh et al., 2023)	Events	General, mixed	Human	115	6,112
RAMS (Ebner et al., 2020)	Events	Politics and conflict, news	Human	139	15,786
PHEE (Sun et al., 2022)	Events	Pharmacovigilance, case reports	Human	2	8,537
CC-News	Events	General, news	LLM (ours)	45,543	295,343

Table 12: Training sources using real texts. *Human* denotes the original human annotations, *LLM (released)* denotes released LLM annotations, and *LLM (ours)* denotes annotations produced by our pipeline. Type counts include names with at least one training annotation, counting relation types for REBEL and event types for event sources. Example counts refer to converted training instances, including trigger and argument instances derived from event passages. Synthetic sources are described in Table 13 and Figure 5.

Genres	email; text message conversation; internal business report; meeting notes; social media post; product review; short news brief; customer support ticket; invoice or order confirmation; job posting; contract clause; clinical note; research paper abstract; recipe; travel itinerary; forum discussion post; press release; resume summary; encyclopedia-style paragraph; podcast transcript excerpt; spoken request to a voice assistant; search query log; legal contract summary; government notice; patent abstract; wiki talk page comment; product specification sheet; hospital discharge summary; sports match report; obituary; museum exhibit label; travel guide paragraph; restaurant menu description; movie synopsis; tweet thread; answer on a question-and-answer site; tutorial step; software changelog entry; science news article; court judgment excerpt
Domains	finance and banking; healthcare; pharmaceuticals; law and courts; politics and elections; government and public policy; military and defense; sports; music; film and television; literature and publishing; visual arts; video games; fashion; food and restaurants; travel and tourism; hotels and hospitality; real estate; construction and architecture; automotive; aviation; shipping and logistics; retail and e-commerce; consumer electronics; telecommunications; software engineering; cybersecurity; artificial intelligence research; data science; physics; chemistry; biology and genetics; astronomy and space; earth science and climate; agriculture; energy and utilities; oil and gas; mining; manufacturing; human resources; education; academia and universities; religion; history; archaeology; philosophy; linguistics; journalism and media; advertising and marketing; insurance; cryptocurrency and blockchain; nonprofit and charity; environmental conservation; veterinary and animals; sports betting and gaming; social work; public transport; weddings and events; fitness and wellness; parenting and family; ophthalmology; dentistry; mental health; nutrition; tennis; association football; basketball; baseball; cricket; motorsport; cycling; the Olympic Games; chess; board games; anime and manga; theatre; dance; photography; sculpture; museums; classical antiquity; medieval history; military history; genealogy; urban planning; maritime industry; railways; space industry; robotics; semiconductors; cloud computing; open-source software; mobile apps; social networks; e-sports; tax law; intellectual property; immigration; criminal justice; human rights; international relations; diplomacy; local government; labor unions; supply chains; wine and beverages; coffee and tea; gardening; pets; zoology; botany; marine biology; ecology; geology; meteorology; oceanography; epidemiology; neuroscience; materials science; mathematics

Table 13: Genres and domains used for synthetic generation. Each request combines one genre with one domain and asks the generator to produce a text and an extraction schema.

Prompt Template (Entities).

Write one realistic {genre} from the domain of {domain}, about {length} words long, in English. Make it specific and natural, with concrete names, places, numbers and dates where the genre calls for them.

Then design an entity-extraction schema for this text: 6 to 12 entity types a user working in {domain} would want to extract from texts like it. Include at least two pairs of closely related types that must be told apart (for example a general type and a more specific one, or two similar roles). Give every type a one-sentence definition that would apply to other texts too. Annotate every mention of every schema type exactly as it appears in the text; schema types that do not occur get no entries. Finally list 3 to 5 further entity types that are relevant to {domain} but do not occur anywhere in this text, each with a one-sentence definition.

Return JSON only: {"text": "...", "types": {"type name": "definition"}, "entities": [{"text": "...", "type": "..."}], "absent_types": {"type name": "definition"}}

Prompt Template (Relations).

Write one realistic {genre} from the domain of {domain}, about {length} words long, in English. Make it specific and natural, with concrete names, places, organisations, products, numbers and dates where the genre calls for them, and with several entities that are related to each other.

Then design an extraction schema for this text:

- 6 to 12 entity types a user working in {domain} would want to extract, each with a one-sentence definition that would apply to other texts too;

- 5 to 10 directed relation types between those entities (from a head entity to a tail entity), each with a one-sentence definition. Include at least two pairs of closely related relation types that must be told apart (for example “founded” and “employed by”, or a relation and its reverse direction).

Annotate every mention of every entity type exactly as it appears in the text, and every relation of a schema type that the text states between two of those mentions (head and tail copied exactly as they appear). Schema types that do not occur get no entries. Finally list 3 to 5 further relation types that would be plausible between these entities but are not stated in the text, each with a one-sentence definition.

Return JSON only: {"text": "...", "entity_types": {"type": "definition"}, "entities": [{"text": "...", "type": "..."}], "relation_types": {"relation": "definition"}, "relations": [{"head": "...", "relation": "...", "tail": "..."}], "absent_relations": {"relation": "definition"}}

Prompt Template (Events).

Write one realistic {genre} from the domain of {domain}, about {length} words long, in English. Make it specific and natural, with concrete names, places, numbers and dates where the genre calls for them, and let it report several events (things that happen or are done), with their participants, times, places and other details.

Then design an event schema for this text: 3 to 6 event types a user working in {domain} would want to extract, each with a one-sentence definition that would apply to other texts too, and for each event type 2 to 6 argument roles (the participants and attributes of such an event), each with a one-sentence definition. Include at least one pair of closely related event types that must be told apart. Annotate every event of every schema type in the text: its type, its trigger (the single word, occasionally a short phrase, that most directly expresses the event, copied exactly) and its arguments (role and the span filling it, copied exactly as it appears; only roles of that event type). Event types of the schema that do not occur get no events. Finally list 2 to 4 further event types that are relevant to {domain} but do not occur anywhere in this text, each with a one-sentence definition.

Return JSON only: {"text": "...", "event_types": {"type": {"definition": "...", "roles": {"role": "definition"}}}, "events": [{"type": "...", "trigger": "...", "arguments": [{"role": "...", "text": "..."}]}], "absent_event_types": {"type": "definition"}}

Figure 5: Generation prompts for entities, relations, and events. The genre and domain values come from Table 13. Requested text lengths are 30, 60, 120, 200, or 300 words. Types listed as absent provide explicit negative supervision after annotation checking.

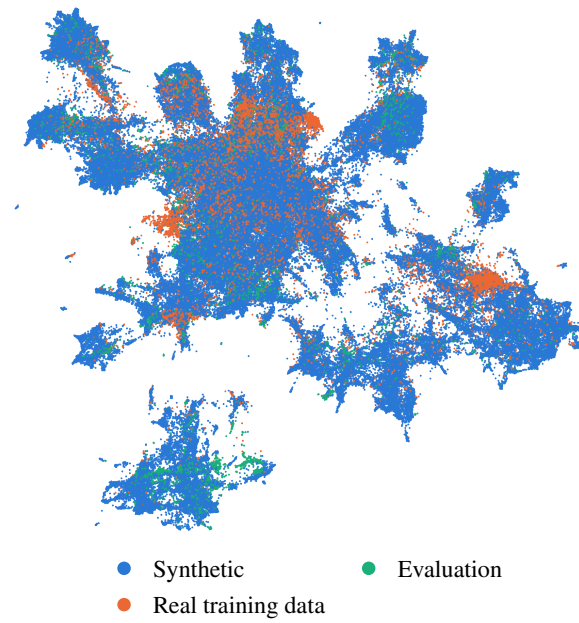


Figure 6: UMAP (McInnes et al., 2018) projection of sentence embeddings sampled from synthetic training sources, real-text training sources, and evaluation datasets.

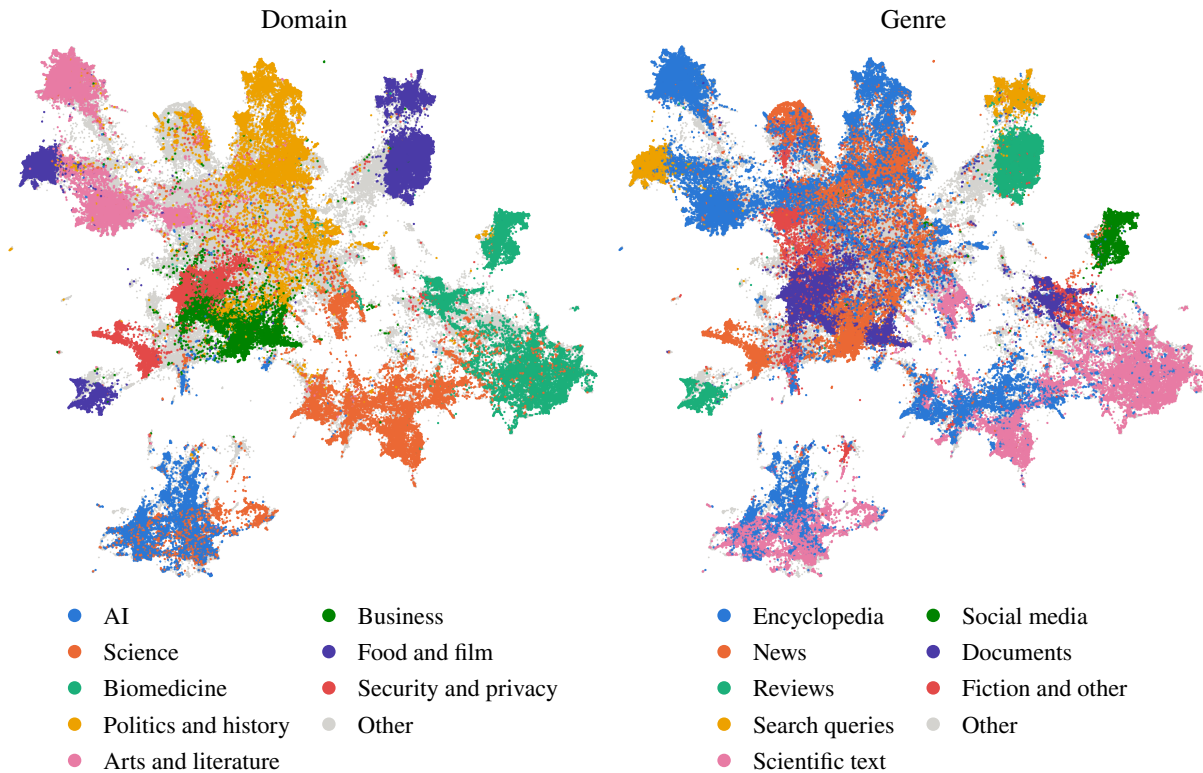


Figure 7: The projection in Figure 6 coloured by domain on the left and genre on the right. Texts outside the listed categories are shown in grey.