

CANE is a Format: Rethinking Output Formats for Few-Shot LLM-Based Named Entity Recognition

Joseph Cornelius^{1,2} Kei K. Ito^{3,4} Fabio Rinaldi¹

¹Dalle Molle Institute for Artificial Intelligence Research (IDSIA), Lugano, Switzerland

²Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

³Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

⁴Broad Institute of Harvard and MIT, Cambridge, MA, USA

{joseph.cornelius, fabio.rinaldi}@idsia.ch kei_ito@dfci.harvard.edu

Abstract

Output format selection is a necessary but understudied dimension of few-shot named entity recognition (NER) with large language models. Existing formats carry characteristic failure modes under autoregressive generation, yet output format has not been systematically evaluated as a primary variable in the few-shot setting across diverse domains and entity structural complexity. We evaluate 21 output formats over 17 datasets across two levels of structural entity complexity, using 9 open-weight models without fine-tuning, ranging from 1.7B to 70B parameters. Output format is the main factor affecting performance, with a 41.7 F₁ points difference between the best and worst formats. This exceeds the gains from scaling models parameters within-family. Furthermore, output format accounts for approximately 49.1% of F₁ variance on simple datasets and 74.9% on complex ones. We describe three mechanistically distinct failure modes that explain these observed differences; each existing format violates at least one of three properties jointly necessary for robust few-shot NER. We propose CANE (Context-Anchored Natural Expression), a format family addressing all three properties. All CANE variants ranked the highest on simple and complex datasets. This includes the GENIA dataset where all other formats had near-zero scores across nine models, indicating the necessity of the three properties. These findings highlight the importance of output format in few-shot named entity recognition, particularly for annotations involving overlapping or discontinuous entities.

1 Introduction

Deploying large language models (LLMs) for named entity recognition (NER) (Grishman and Sundheim, 1996; Chinchor and Robinson, 1998) requires choosing an output format (how the model should express its predictions as text). The literature offers a wide range of formats originating from

research in annotation tooling, structured prediction, and LLM prompting (Ramshaw and Marcus, 1995; Stenetorp et al., 2012; Wang et al., 2025; Ren and Teo, 2025; Zhan et al., 2026). Yet practitioners choose among them on the basis of familiarity rather than empirical evidence. While recent work demonstrates substantial format effects in fine-tuned settings (Ravikumar et al., 2026; Zhan et al., 2026), the few-shot setting remains understudied across various domains, model families, and entity structural complexity. Generative LLMs allow arbitrary output formats, yet each existing format carries a failure mode under LLM inference. Offset-based and inline markup formats impose requirements that autoregressive generation cannot reliably meet, which includes exact character counting and verbatim input copying. Per-token formats require positional label-token correspondence while entity enumeration formats cannot resolve repeated surface forms or represent overlapping and discontinuous entities (Raman et al., 2022; Wang et al., 2025; He and Choi, 2023). These failure modes define three properties a format must satisfy simultaneously: a surface form aligned with training distributions, unambiguous span grounding without character counting or verbatim copying, and native expressivity for complex entity structures. No existing format satisfies all three.

We present an evaluation of 21 NER output formats across 17 datasets and two levels of entity structural complexity using 9 open-weight models from 1.7B to 70B parameters, under five-shot prompting without fine-tuning. Our experiments show that format choice is the main factor affecting model performance, with the difference in F1 score between best and worst performing format being 41.7 F1 points. This exceeds the gain from scaling within any single model family.

We demonstrate that format choice is the main determinant in few-shot LLM-based NER, consistent across model families and domains, and

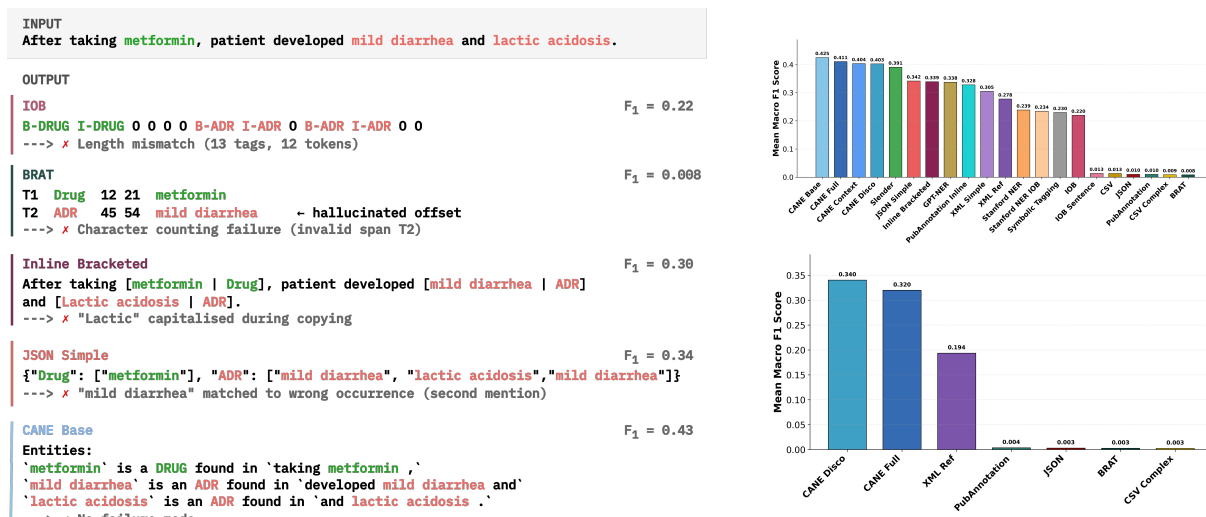


Figure 1: Format families and their effect on few-shot NER performance. Left: representative outputs illustrating each format group and its failure mode. Right: mean F_1 across all evaluated formats on simple (top) and complex (bottom) datasets, colour-coded by format group. The 41.7-point gap between best and worst formats on simple datasets, and the collapse of all non-CANE formats to 0.000 on GENIA, establish format choice as the dominant performance variable.

that existing formats each violate at least one of the three jointly necessary properties above. We then propose CANE (Context-Anchored Natural Expression), an output format that extends the natural-language enumeration schema of He and Choi (2023), by introducing context anchoring and explicit structural encoding of entity complexity, without modifying the inference procedure (Figure 1). Finally, we showcase three distinct failure modes explaining the format hierarchy with an error taxonomy, and an ablation study isolates the contribution of different CANE design features.¹

Sections 2–4 position the work and establish the evaluation design; Sections 5–7 deliver the experimental argument; Section 9 interprets findings.

2 Related Work

NER output formats. Token-level sequence formats, e.g., IOB (Ramshaw and Marcus, 1995) and its variants IOB2 (Ratnaparkhi, 1998) and BILOU (Ratinov and Roth, 2009), remain standard for discriminative sequence labellers and LM-based token classifiers (Panchendrarajan and Amaesan, 2018; Devlin et al., 2019). Offset-based formats, including BRAT Standoff (Stenetorp et al., 2012), PUBANNOTATION JSON (Kim and Wang, 2012), and BIONLP shared task representations (Kim et al., 2011), prevail in biomedical

corpus construction, supporting the full structural range but requiring exact character indexing. Inline markup formats such as INLINE XML (Keloth et al., 2024) embed annotation within the input string; GPT-NER (Wang et al., 2025) adapts this for LLM generation, later extended to multi-type annotation (Yan et al., 2025). Natural-language enumeration formats list entities as structured prose (Cui et al., 2021; He and Choi, 2023); SLENDER (Ren and Teo, 2025) preserves performance while reducing inference costs on flat benchmarks. Semin et al. (2026) defined a tagging/ indexing/ matching taxonomy and found that indexing is unreliable without explicit positional markers, corroborating the offset-based failure mode in Section 1.

Closest prior work. Cui et al. (2021) scores pre-enumerated candidate spans against a prose template using BART (Lewis et al., 2020), requiring exhaustive candidate generation. He and Choi (2023) extend this to one-pass generation with PT, showing natural-language schemas are easier for fine-tuned sequence-to-sequence models to learn than label-sequence or markup linearizations. However, they acknowledged two limitations: string matching is noisy when an entity surface form appears in non-entity context, and nested and discontinuous entities are excluded. CANE addresses both limitations within LLM few-shot setting.

¹Code is available at: <https://github.com/IDSIA-NLP/cane>

Dataset	Avg Doc Length (Tokens)	Category	Complex	Num Labels	Num N O D	Entity Density /Doc	Avg Unique Labels /Doc	Num Test Docs
CONLL2003	13.21	News	✗	4	0 1 0 1 0	1.64	1.19	1000
NCBI DISEASE	26.03	Medical	✗	1	0 1 0 1 0	1.02	0.57	941
BIONLP2004	26.07	Biomedical	✗	5	0 1 0 1 0	2.25	1.28	1000
WNUT2017	18.45	Social Media	✗	6	0 1 0 1 0	0.86	0.68	1000
BC5CDR	19.77	Biomedical	✗	2	0 1 0 1 0	1.73	1.03	1000
FIN	43.44	Financial	✗	4	0 1 0 1 0	1.04	0.44	305
ONTONOTES5	18.38	General	✗	18	0 1 0 1 0	1.35	1.02	1000
WIKINEURAL	22.92	General	✗	16	0 1 0 1 0	1.61	1.19	1000
MULTINERD	21.95	General	✗	17	0 1 0 1 0	1.6	1.09	1000
WIKIANN	8.12	General	✗	3	0 1 0 1 0	1.43	1.16	1000
WIKIANN[DE]	9.64	General	✗	3	0 1 0 1 0	1.39	1.07	1000
WIKIANN[FR]	6.88	General	✗	3	0 1 0 1 0	1.32	1.06	1000
WIKIANN[ZH]	22.72	General	✗	3	0 1 0 1 0	1.25	1.02	1000
CADEC	84.04	Medical	✓	1	728 1 43 1 798	5.11	0.88	1000
SHARE2013 P	36.26	Medical	✓	1	134 1 2 1 185	2.21	0.46	1000
SHARE2014 P	41.17	Medical	✓	1	124 1 41 1 171	2.23	0.46	1000
GENIA	216.73	Biomedical	✓	39	17859 1 0 1 0	48.59	8.52	1000

Table 1: Dataset statistics. **Complex**: whether the corpus contains nested (N) overlapping (O), or discontinuous (D) entities. **N|O|D**: entity counts per split. **Entity Density**: mean entities per document. **Avg Unique Labels**: mean distinct entity types per document.

Format sensitivity and few-shot NER. Ravikumar et al. (2026) show that format choice alone shifts F_1 by more than 40 points across fine-tuned NER, event detection, and semantic role labelling; Zhan et al. (2026) confirm that offset-based formats collapse to near-zero F_1 in fine-tuned NER and that the optimal format depends on annotation characteristics. More broadly, delimiter type, separator, and schema phrasing have measurable effects on LLM output quality (Sclar et al., 2024; Su et al., 2025; Long et al., 2025). Semin et al. (2026) address span grounding via logit-level constrained decoding, inapplicable to API-accessed models.

Few-shot NER and structural complexity. Instruction-tuned LLMs achieve competitive few-shot NER performance (Wang et al., 2025; Zhou et al., 2024; Sainz et al., 2024), but performance on structurally complex annotations, i.e., nested, overlapping, and discontinuous entities prevalent in biomedical and clinical text (Karimi et al., 2015; Mowery, 2013a), is substantially lower across model scales (Kim et al., 2024; Cabral et al., 2025; Alhassan et al., 2025). Prior format comparisons have not varied structural entity complexity as a controlled dimension (Ravikumar et al., 2026; Zhan et al., 2026).

Therefore, we study the impact of structurally complex entities on format rankings and whether any existing format can express the full structural range. To our knowledge, no prior study evaluates output format as a primary independent variable in few-shot NER across diverse domains and both flat and structurally complex entity annotation.

3 Datasets

We evaluate 17 datasets spanning five domains and two levels of entity structural complexity; statistics and citations are in Table 1 and Table 5.

Entity structures. We follow Dai (2018) and Wang et al. (2022) in the definition of entity structural complexity; examples are given in Fig. 7.

Flat: an entity spanning contiguous tokens that shares no token with any other entity.

Nested: an entity whose token span is properly contained within that of another entity.

Overlapping: two entities sharing at least one token, with neither span contained in the other.

Discontinuous: an entity whose tokens are separated by at least one other intervening token.

Simple datasets. 13 corpora contain only flat entities, covering news, general knowledge, social media, financial, and biomedical text.

Complex datasets. 4 corpora contain overlapping or discontinuous entities. CADEC is a medical patient-forum corpus with 728 nested, 42 overlapping and 798 discontinuous adverse drug reaction mentions. SHARE-2013 and SHARE-2014 contain discontinuous and nested disorder mentions. They are segmented at paragraph boundaries into sub-documents, resulting in SHARE-2013-P, SHARE-2014-P, due to document length constraints. GENIA has documents more than twice as long on average as the next-longest dataset, 39 entity types, and 17,859 nested entity pairs.

Evaluation protocol. We sample 1,000 documents per dataset from the official test split (or use the full split when fewer are available), and map corpus-specific labels to human-readable semantic names prior to prompting (e.g., B-PER → PERSON).

4 Formats

4.1 Existing Formats

Producing structured NER output from a generative LLM requires identifying spans and labeling their types as free-form text, a challenge each existing format family resolves differently, and none resolves without a characteristic failure mode in autoregressive generation. Three properties (P) are necessary for robust NER for LLM inference: (i) a surface form aligned with pretraining distribution, (ii) unambiguous span grounding without verbatim copying or character counting, and (iii) native

Format	Example	F	N	O	D
IOB	After 0\nstarting 0\namoxicillin B-DRUG\n, 0\nthe 0\npatient 0\ndeveloped 0\nna 0\nrash B-ADR\n. 0	✓	✗	✗	✗
IOB SENTENCE	0 0 B-DRUG 0 0 0 0 B-ADR 0	✓	✗	✗	✗
STANFORD NER IOB	After/0 starting/0 amoxicillin/B-DRUG ,/0 the/0 patient/0 developed/0 a/0 rash/B-ADR ./0	✓	✗	✗	✗
STANFORD NER	After/0 starting/0 amoxicillin/DRUG ,/0 the/0 patient/0 developed/0 a/0 rash/ADR ./0	✓	✗	✗	✗
CSV	0,0,B-DRUG,0,0,0,0,0,B-ADR,0	✓	✗	✗	✗
PUBANNOT. INLINE	After starting [amoxicillin][DRUG], the patient developed a [rash][ADR].	✓	✓	✗	✗
SYMBOLIC TAGGING	DRUG: After starting [amoxicillin], the patient developed a rash.\nADR: After starting amoxicillin, the patient developed a [rash].	✓	✓	✗	✗
INLINE BRACKETED	After starting [amoxicillin DRUG], the patient developed a [rash ADR].	✓	✓	✗	✗
GPT-NER	After starting @@amoxicillin##DRUG, the patient developed a @@rash##ADR.	✓	✓	✗	✗
XML SIMPLE	After starting <DRUG>amoxicillin</DRUG>, the patient developed a <ADR>rash</ADR>.	✓	✓	✗	✗
JSON SIMPLE	{"DRUG": ["amoxicillin"], "ADR": ["rash"]}	✓	✓	✓	✗
SLENDER	amoxicillin DRUG; rash ADR	✓	✓	✓	✗
CANE-BASE	Entities:\n'amoxicillin' is a DRUG\n'rash' is an ADR	✓	✓	✓	✓
CANE-CONTEXT	Entities:\n'amoxicillin' is a DRUG found in 'starting amoxicillin,' \n'rash' is an ADR found in 'a rash.'	✓	✓	✓	✗
XML REF	<DISEASE ref="e1">metastatic <CELL ref="e2">breast cancer</DISEASE> cells</CELL> caused <ADR ref="e3">headache</ADR> and <ADR ref="e4">nausea</ADR>	✓	✓	✗	✓
PUBANNOTATION	{ "id": "T1", "span": { "begin": 0, "end": 24 }, "obj": "DISEASE" }, { "id": "T2", "span": { "begin": 11, "end": 30 }, "obj": "CELL" }, { "id": "T3", "span": { "begin": 38, "end": 46 }, "obj": "ADR" }, { "id": "T4", "span": { "begin": 51, "end": 57 }, "obj": "ADR" }	✓	✓	✓	✗
CSV-COMPLEX	e1,DISEASE,0:24,...\ne2,CELL,11:30,...\ne3,ADR,38:46;51:57,...	✓	✓	✓	✓
JSON	[{"label": "DISEASE", "spans": [{"start": 0, "end": 24}]}, {"label": "CELL", "spans": [{"start": 11, "end": 30}]}, {"label": "ADR", "spans": [{"start": 38, "end": 46}, {"start": 51, "end": 57}]]\n]	✓	✓	✓	✓
BRAT	T1\tDISEASE 0 24\tmetastatic breast cancer\nT2\tCELL 11 30\tbreast cancer cells\nT3\tADR 38 46;51 57\theadache nausea	✓	✓	✓	✓
CANE-DISCO	Entities:\n'metastatic breast cancer' is a DISEASE\n'breast cancer cells' is a CELL\n'headache' and 'nausea' are an ADR	✓	✓	✓	✓
CANE-FULL	Entities:\n'metastatic breast cancer' is a DISEASE found in 'metastatic breast cancer cells'\n'breast cancer cells' is a CELL found in 'metastaticbreast cancer cells caused'\n'headache' and 'nausea' are an ADR found in 'caused headache and' and 'and nausea.'	✓	✓	✓	✓

Table 2: NER output formats and their structural capabilities (✓ = supported, ✗ = not supported). *Simple* formats are illustrated with “After starting amoxicillin, the patient developed a rash” (DRUG: amoxicillin, ADR: rash). *Complex* formats use “metastatic breast cancer cells caused headache and nausea” (DISEASE: metastatic breast cancer, CELL: breast cancer cells [overlapping with DISEASE], ADR: headache + nausea [discontinuous]). F=Flat, N=Nested, O=Overlapping, D=Discontinuous.

support for overlapping and discontinuous entities. We organize existing approaches into four groups; structural capabilities are summarized in Table 2. All four groups comprise established NER annotation formats and are evaluated under identical conditions.

Per-Token Label Sequences (Group 1). Formats such as IOB2 (Ratnaparkhi, 1998), IOB SENTENCE (tag sequence alone, without tokens), CSV, and STANFORD NER token-label pairs (Finkel et al., 2005) emit one label per input token, making span location unambiguous by construction. However, they require strict positional correspondence between every generated tag and its input token (failure of P(i)). This assumption is violated by autoregressive generation (Wang et al., 2025; Rahman et al., 2022), and they cannot represent nested, overlapping, or discontinuous entities (failure of P(iii)).

Inline Markup (Group 2). GPT-NER (Wang et al., 2025; Yan et al., 2025), PUBANNOTATION INLINE (Kim and Wang, 2012; Kim, 2026), SYMBOLIC TAGGING (Lv et al., 2025) and XML SIMPLE (Keloth et al., 2024) formats embed entity boundaries within a verbatim copy of the input, presupposing exact reproduction. However, LLMs

introduce alterations (whitespace normalization, punctuation modification) during generation that corrupt span boundaries (Jarolím et al., 2025; Haque et al., 2026), which is exacerbated by document length (Du et al., 2025; Liu et al., 2024) (failure of P(ii)).

Offset-Based Structured Output (Group 3). BRAT (Stenetorp et al., 2012), PUBANNOTATION (Kim and Wang, 2012), and their JSON and CSV COMPLEX counterparts identify spans by character-level indices, which LLMs consistently fail at (Zhang and He, 2024; Shin and Kaneko, 2024) (failure of P(ii)).

Entity Enumeration (Group 4). JSON SIMPLE (Zhan et al., 2026), PT FORMAT (He and Choi, 2023) and SLENDER (Ren and Teo, 2025) list entity strings paired with types, resolving span location by post-hoc string matching. This is the most aligned with autoregressive text generation, but with two limitations: repeated surface forms make matching ambiguous (partial failure of P(ii)), and no mechanism exists for discontinuous entities (failure of P(iii)).

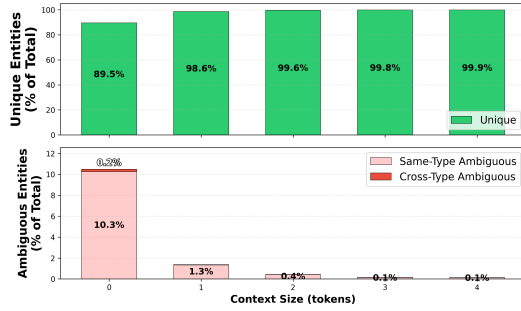


Figure 2: Proportion of entity surface forms disambiguated by bidirectional context anchoring of increasing size (tokens), aggregated across all 17 datasets with equal weighting. A surface form is disambiguated if its context window uniquely identifies its occurrence.

4.2 Proposed Format

No existing family simultaneously satisfies all three properties listed above. Group 1 fails P(i) and P(iii); Groups 2 and 3 fail P(ii); Group 4 satisfies P(i) but fails P(iii) and partially P(ii) for repeated surface forms. While P(i) relates directly to performance, P(ii) and P(iii) concern a format’s ability to annotate spans unambiguously and to represent complex structures, respectively, so satisfying them is a precondition for handling ambiguous and complex annotation and does not by itself imply higher F_1 . We propose **Context-Anchored Natural Expression** (CANE), a format family extending natural-language enumeration with context anchoring and native structural expressivity.

Design Rationale. CANE is based on three principles. First, output should resemble structured prose present in pretraining and instruction-tuning corpora; formats departing from the training distribution impose a performance cost apparent under few-shot inference (Tam et al., 2024; Lee et al., 2026). Second, each entity is identified by a minimal verbatim context window from the input via *found in*, eliminating both the copying failure of inline markup and the counting failure of offset-based formats while resolving occurrence ambiguity: Figure 2 shows that one-token bidirectional context reduces ambiguous surface forms from 10.5% to 1.4%, reaching 99.9% unique identification at four tokens across all 17 datasets. Third, overlapping entities are generated on independent lines that share a context anchor; discontinuous entities use and-conjunction with one anchor per fragment, both without schema change.

Format Specification. The **primer** (Entities:) frames output as a list; **entity** spans in backticks identify the surface form; **types** appear as upper-case labels; **context anchors** provide a verbatim context resolving ambiguity without character counting. Token-level choices, the selection of individual tokens that make up the format, are evaluated in Section 8. The full structural capacity handles complex entities without schema change:

```
Entities:\n
'entity' is a TYPE found in 'context_left entity context_right'\n
'entity_1' and 'entity_2' are an TYPE found in 'context_left entity_1\n
context_right' and 'context_left entity_2 context_right'\n
```

Examples are showcased in Fig. 1 and Table 2.

Format Variants. We define a 2×2 family over context anchoring and discontinuous encoding. CANE-BASE omits both, producing *'entity' is a TYPE* per entity, and satisfies P(i); CANE-CONTEXT adds anchoring (P(ii)); CANE-DISCO adds discontinuous annotation (P(iii)); CANE-FULL retains both and satisfies all three properties. CANE-BASE is structurally equivalent to the PT schema (He and Choi, 2023) with differences in token-level choices outlined in Section 8. It is optimized for the few-shot setting and outperforms PT by 3.1 F_1 points. The 2×2 design makes each feature’s cost and benefit interpretable, expected to vary with dataset complexity and model scale.

5 Methods

We evaluate different output formats on various models and datasets. All other conditions, including prompt structure, in-context examples, decoding, and evaluation, are held constant. No model is fine-tuned.

Models. We evaluate nine open-weight instruction-tuned LLMs spanning 1.7B to 70B parameters across four families: Qwen3 (1.7B, 8B, 30B-A3B (Yang et al., 2025)), LLaMA-3 (3.2-3B, 3.3-70B (Grattafiori et al., 2024)), Mistral (7B-v0.3, Small-24B, 8x7B (Team, 2025; Jiang et al., 2023, 2024)), and Gemma-3 (4B (Gemma Team et al., 2025)), all served via vLLM (Kwon et al., 2023). An ablation study (Section F.1) evaluates Qwen3 (1.7B, 8B) with reasoning enabled. Details of decoding parameters and computational budget are in Appendix C.

Prompting setup. Each prediction uses a five-shot prompt comprising a task description, a format specification, a constraint against altering the input,

and five randomly sampled examples held constant across predictions. For complex datasets, examples include at least one discontinuous and nested case. We use a one-token bidirectional context window for all CANE variants.

Parsing and post-processing. Each output is converted to a unified representation by a format-specific parser; details are in Appendix D. We treat the parser as part of the format pipeline, since in deployment a format is only usable together with a parser. Two error conditions yield an empty prediction list: *malformed output* and *length mismatch* (Group 1 only). Two others discard only the affected entity: *invalid label* and *span mismatch*. For CANE, context-anchored variants (CANE-FULL, CANE-CONTEXT) use two-stage anchor-then-span matching with per-entity failure handling, while non-anchored variants (CANE-BASE, CANE-DISCO) use direct first-match resolution with document-level failure detection. The properties are format-intrinsic, while parser strictness is an implementation choice; parsers were designed with comparability as the primary criterion. This dependency is examined in Section 7 and, for the CANE variants, in Appendix H.

Evaluation metric. We report strict macro F_1 , counting a prediction correct only when both span boundaries and entity type exactly match the annotation; letting P_c and R_c denote precision and recall for type $c \in C$:

$$F_1^{\text{macro}} = \frac{1}{|C|} \sum_{c \in C} \frac{2P_c R_c}{P_c + R_c}. \quad (1)$$

Macro F_1 prevents formats from being rewarded on corpora dominated by a single entity type. F_1 scores are reported on a 0–1 scale; differences are expressed in percentage points (multiplied by 100).

6 Results

6.1 Format Choice

Format choice determines few-shot NER performance across all evaluated models and datasets. The gap between the best-performing format (CANE-BASE, mean $F_1 = 0.425$) and the worst (BRAT, 0.008) is 41.7 F_1 points averaged across 10 simple datasets and 9 models (Table 4, Figure 1). Format accounts for 49.1% of F_1 variance on simple and 74.9% on complex datasets, exceeding the contribution of model and dataset in both conditions (Figure 4). These proportions derive from

a group-means decomposition and indicate relative magnitude rather than an exact variance partition. Results fall into three tiers: a top tier of all four CANE variants and SLENDER (0.391–0.425); a middle tier of enumeration and inline markup formats including IOB (0.220–0.342); and a bottom tier of offset-based and label-only per-token formats that collapses to near-zero F_1 , no format exceeding 0.041 on any model–dataset combination.

The difference in performance across annotation groups reflects failures that precede parsing. The offset-based formats might have a valid structure, but the indices do not match the input span, while per-token formats break the label–token correspondence their parsers require, leaving no annotation to recover in either case. Parser strictness shifts the ranking within the CANE family (Section H), but by a margin smaller than the separation between families.

Section 7 shows this collapse reflects an incompatibility with autoregressive generation.

6.2 Simple Datasets: CANE Family Leads

All four CANE variants rank in the top five, with CANE-BASE best (0.425), followed by CANE-FULL (0.411), CANE-CONTEXT (0.404), and CANE-DISCO (0.403). The closest non-CANE format, SLENDER (0.391), trails CANE-BASE by 3.4 points. The within-family ordering is consistent with the design: context anchoring and discontinuous encoding add overhead with no benefit on flat unambiguous entities, so CANE-BASE leads and CANE-DISCO trails by 2.2 points (Section 6.3). The sole exception is Mistral-7B-v0.3, where CANE-FULL outperforms CANE-BASE by 12.1 points (0.385 vs 0.264), reflecting sensitivity to context anchoring syntax (Appendix F.6). JSON SIMPLE is malformed in 24% of predictions, resulting in an 8.3-point gap with CANE-BASE. CANE-DISCO outperforms CANE-FULL by 2.4 F_1 points when controlled for parser success (Appendix H).

6.3 Complex Datasets: Structural Expressivity Is Necessary

CANE-DISCO achieves the highest mean F_1 across all four complex corpora (0.340), outperforming CANE-FULL on all four datasets and eight of nine models (Figure 3). The gap to XML REF, the only non-CANE format with structural expressivity, is between 12.8 and 16.4 points on CADEC and SHARE-2013/ 2014-P. On GENIA, every non-

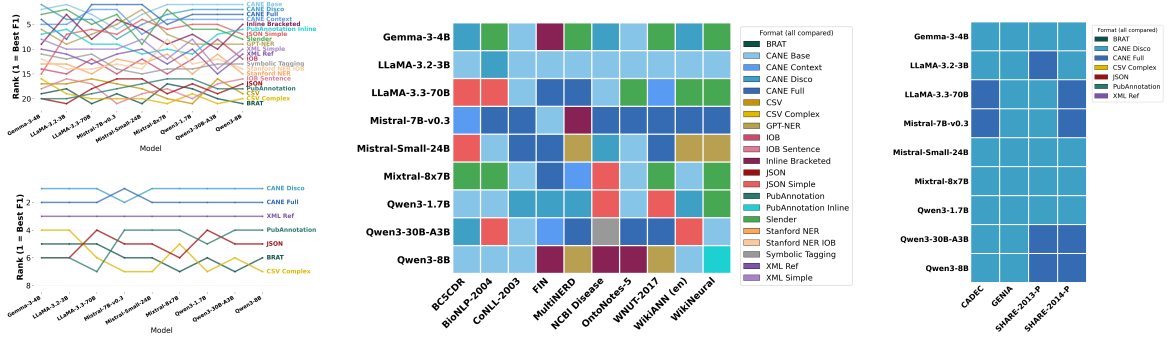


Figure 3: Format ranking consistency across models and datasets. Left: format rank across 9 models on simple (top) and complex (bottom) datasets; lower rank is better. Centre and right: best-performing format per model $\times$ dataset combination on simple and complex datasets respectively.

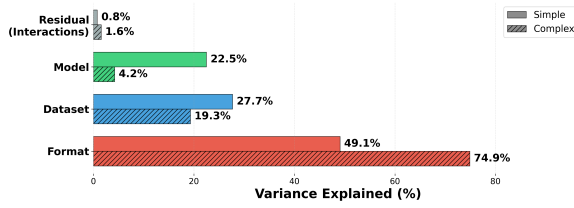


Figure 4: Proportion of strict macro F_1 variance attributable to each factor (variance of group means approximation).

Format	CADEC	GENIA	SHARE-2013-P	SHARE-2014-P	Avg
CANE-DISCO	0.458	0.165	0.342	0.396	0.340
CANE-FULL	0.456	0.102	0.335	0.387	0.320
XML REF	<u>0.294</u>	0.000	<u>0.214</u>	<u>0.267</u>	<u>0.194</u>
PUBANNOTATION	0.009	0.000	0.003	0.004	0.004
JSON	0.008	0.000	0.002	0.003	0.003
BRAT	0.007	0.000	0.002	0.002	0.003
CSV COMPLEX	0.007	0.000	0.002	0.002	0.003

Table 3: Strict macro F_1 on four complex datasets, averaged across 9 models. **Bold**: best per dataset. Underline: best non-CANE score.

CANE format scores at most 0.001 across all nine models; CANE-DISCO achieves 0.165 and CANE-FULL 0.102.

6.4 Format Rankings Are Consistent Across Model Families

Format group rankings are stable across all four model families (Figure 3): CANE variants consistently occupy the top ranks and offset-based formats the bottom, with a mean Kendall rank correlation of $\tau = 0.726$ (range 0.587–0.838) across model pairs on simple datasets. Absolute F_1 varies from 0.264 (Mistral-7B-v0.3) to 0.560 (LLaMA-3.3-70B) for CANE-BASE. Offset-based formats fail regardless of model scale: LLaMA-3.3-70B scored 0.018, matching Qwen3-1.7B’s 0.002 on simple datasets with BRAT.

7 Error Analysis

The format hierarchy reflects three mechanistically distinct failure modes; Figure 5a shows aggregate error rates on simple datasets (Span Error per entity therefore can exceed 1.0). The errors are categorized as malformed (structurally unparseable predictions); span error (span-matching failures per prediction); labeling error (wrong entity type), analogous to (Semin et al., 2026).

Critical errors. Group 1 fails because autoregressive generation cannot maintain positional label-token correspondence: IOB SENTENCE and CSV produce length mismatches on 0.93 and 0.94 of predictions; IOB is less extreme (0.40), consistent with its F_1 of 0.220. Offset-based formats split into two sub-populations: JSON and PUBANNOTATION has malformed output on 0.29 of predictions; BRAT and CSV COMPLEX produce parseable output but hallucinate character offsets on almost every entity (span errors 1.36 and 1.03).

Non-critical span errors. Group 2 fails because LLMs alter input during copying slightly, corrupting span boundaries even for correctly identified entities (INLINE BRACKETED span error 0.41, XML REF 0.29). Span errors in CANE variants reflect failure propagation differences, reaching 0.41 in context-anchored versus 0.17 in non-anchored variants.

Complex datasets and the GENIA mechanism. On longer, denser complex datasets (Figure 5b), JSON and PUBANNOTATION reach malformed rates near 0.50. GENIA’s complexity increases span error to 4.03 per prediction for CANE-DISCO, versus 0.47 on CADEC. Span-matching exhaustion follows from dense nested (repeated surface forms),

Format	BC5CDR	BioNLP-2004	CoNLL-2003	FIN	MultiNERD	NCBI Disease	OntoNotes-5	WNUT-2017	WikiANN (en)	WikiNeural	Avg
CANE-BASE	0.585	0.326	0.590	0.240	0.328	0.453	0.304	0.309	0.558	0.556	0.425
CANE-FULL	0.542	0.313	0.576	0.241	0.325	0.419	0.291	0.317	0.536	0.545	0.411
CANE-CONTEXT	0.529	0.297	0.568	0.236	0.311	0.418	0.293	0.313	0.538	0.533	0.404
CANE-DISCO	0.565	0.324	0.560	0.224	0.294	0.434	0.291	0.285	0.530	0.519	0.403
SLENDER	<u>0.536</u>	<u>0.301</u>	<u>0.549</u>	0.185	<u>0.319</u>	0.389	<u>0.267</u>	<u>0.298</u>	<u>0.515</u>	<u>0.548</u>	<u>0.391</u>
JSON SIMPLE	0.491	0.253	0.444	0.194	0.246	<u>0.407</u>	0.216	0.264	0.458	0.442	0.342
INLINE BRACKETED	0.434	0.251	0.471	0.194	0.299	0.350	0.232	0.225	0.437	0.499	0.339
GPT-NER	0.432	0.220	0.493	<u>0.207</u>	0.302	0.307	0.224	0.234	0.445	0.513	0.338
PUBANNOTATION INLINE	0.408	0.253	0.460	0.179	0.298	0.310	0.209	0.234	0.417	0.513	0.328
XML SIMPLE	0.393	0.230	0.384	0.179	0.273	0.343	0.179	0.205	0.407	0.456	0.305
XML REF	0.317	0.239	0.347	0.127	0.265	0.310	0.170	0.188	0.375	0.441	0.278
STANFORD NER	0.333	0.140	0.366	0.114	0.169	0.249	0.172	0.154	0.296	0.393	0.239
STANFORD NER IOB	0.302	0.145	0.359	0.104	0.192	0.206	0.171	0.163	0.305	0.398	0.234
SYMBOLIC TAGGING	0.324	0.181	0.272	0.115	0.195	0.325	0.141	0.132	0.280	0.330	0.230
IOB	0.310	0.113	0.352	0.114	0.162	0.176	0.140	0.182	0.295	0.355	0.220
IOB SENTENCE	0.012	0.001	0.033	0.000	0.004	0.003	0.002	0.012	0.053	0.013	0.013
CSV	0.012	0.002	0.026	0.000	0.007	0.002	0.003	0.010	0.056	0.014	0.013
JSON	0.013	0.003	0.035	0.002	0.002	0.002	0.009	0.010	0.014	0.013	0.010
PUBANNOTATION	0.011	0.002	0.037	0.001	0.002	0.002	0.007	0.011	0.016	0.011	0.010
CSV COMPLEX	0.011	0.001	0.034	0.000	0.002	0.002	0.008	0.007	0.013	0.008	0.009
BRAT	0.010	0.002	0.032	0.001	0.002	0.001	0.008	0.006	0.015	0.007	0.008

Table 4: Strict macro F_1 on 10 simple datasets, averaged across 9 models. **Bold**: best per dataset. Underline: best non-CANE score. Formats grouped: CANE variants (top), enumeration, per-token sequence, offset-based (bottom).

too many (39) entity types (type-ambiguous surface forms), and long documents (more matching attempts per prediction). Label errors are small and consistent (0.028–0.115 on simple datasets), confirming demonstrations supply sufficient label vocabulary regardless of format. Representative outputs are in Appendix I.

CSV	0.94	0.00	0.00
IOB Sentence	0.93	0.00	0.00
IOB	0.40	0.00	0.05
PubAnnotation	0.29	0.04	0.05
JSON	0.29	0.03	0.03
Stanford NER IOB	0.24	0.02	0.04
JSON Simple	0.24	0.16	0.03
CANE Disco	0.21	0.17	0.11
Stanford NER	0.18	0.02	0.16
CANE Base	0.13	0.13	0.07
PubAnnotation Inline	0.10	0.35	0.08
CSV Complex	0.10	0.03	0.10
BRAT	0.09	1.36	0.17
Slender	0.05	0.22	0.10
Inline Bracketed	0.04	0.41	0.11
GPT-NER	0.02	0.24	0.12
XML Ref	0.02	0.29	0.12
XML Simple	0.02	0.28	0.14
CANE Context	0.00	0.38	0.10
CANE Full	0.00	0.41	0.10
Symbolic Tagging	0.00	0.20	0.14

(a) Simple Datasets

PubAnnotation	0.59	0.04	0.00
JSON	0.56	0.04	0.01
CSV Complex	0.28	1.60	0.07
BRAT	0.27	1.66	0.03
CANE Disco	0.18	1.31	0.17
XML Ref	0.14	0.33	0.11
CANE Full	0.06	0.92	0.07

(b) Complex Datasets

Figure 5: Error rates per format averaged across all models and datasets. *Malformed*: structurally unparseable predictions (bounded $[0, 1]$). *Span Error*: span-matching failures per prediction (can exceed 1.0). *Labeling Error*: valid span, wrong entity type. Formats sorted by Malformed rate descending.

8 Format Design Analysis

We isolate token-level choices in CANE-BASE via ablations against the PT baseline (He and Choi, 2023); see Figure 6.

These choices are distinct from the properties: the properties determine what a format can represent, the token-level choices affect how reliably a model produces the format. The analysis therefore asks which token-level choices matter, not which format is optimal. As the differences between variants are small relative to the dispersion across

datasets (Table 13), we assess each token choice by its consistency across models and datasets.

Separator Semicolons yield a higher mean F_1 (0.438 vs 0.424), but the gain is driven entirely by Mistral-7B-v0.3 (+0.143); six of nine models prefer newlines (+0.007). Semicolons reduce malformed output (0.09 vs 0.13) but increase span errors (0.60 vs 0.14), trading parse reliability for span grounding accuracy. We adopt newlines for their more predictable error profile.

Span delimiters Replacing backtick with double-quote delimiters reduces F_1 by 0.4 points (0.420 vs 0.424), suggesting backtick highlighting is helpful but not critical.

Primer header Removing the **Entities:** primer is the most consequential change: F_1 drops 4.1 points (0.383 vs 0.424) consistently across all 10 datasets (range: 2.8–6.6 points), with no dataset showing improvement.

PT baseline CANE-BASE outperforms PT Format by 3.1 points. PT Format has low malformed rates (0.04) but high span errors (0.78); backtick delimiters and the primer reduce span matching failures.

9 Discussion

Output format is the main determinant of F_1 score. We observe three failure modes related to format choice (Section 7). We interpret each failure mode as arising from a constraint of autoregressive generation: per-token formats require positional label-token correspondence that decoding cannot maintain; inline markup requires verbatim input reproduction that models alter in practice;

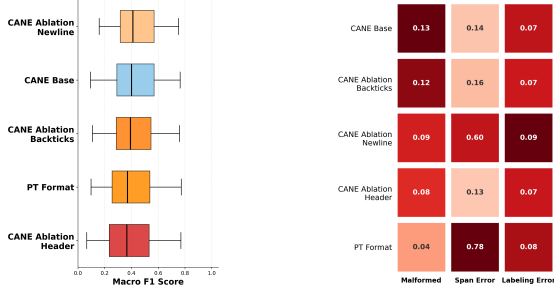


Figure 6: Token-level ablation results on simple datasets (9 models, 10 datasets). Left: strict macro F_1 distribution per variant, sorted by median; Right: error rates per variant. *Malformed*: structurally unparseable predictions. *Span Error*: span-matching failures per prediction (can exceed 1.0). *Labeling Error*: wrong entity type.

and offset-based formats require character-level indexing that models consistently fail (Zhang and He, 2024; Shin and Kaneko, 2024). The hierarchy across four model families (Kendall $\tau = 0.726$) confirms a structural effect, not an artifact of any particular architecture or scale.

The role of annotation complexity. CANE-BASE performs well across all model families on flat, non-overlapping annotation and requires no schema change, as the absence of context anchors and and-conjunction reduces matching errors. CANE-DISCO performs best with complex entities: the 2.2-point cost on simple datasets is offset by a 14.6-point gain over the best alternative on complex corpora, consistent with the error profile showing that context anchors amplify matching complexity, particularly on GENIA (Section 7). Modest scores on GENIA reflect dataset complexity and confirm that all three format properties are necessary for highly complex annotation.

Performance and expressivity are separable. SLENDER and CANE-BASE reach the top tier on flat corpora without P(ii) or P(iii), as expected: 89.5% of occurrences are uniquely annotated without anchoring (Section J), so anchoring add more costs more than it returns, and CANE-DISCO accordingly outperforms CANE-FULL on complex corpora. The higher span-error rate of the anchored variants thus reflects the cost of one mechanism, not a failure of the property. All the CANE variants nonetheless retain their average gains on flat corpora over SLENDER, and on GENIA every P(iii)-satisfying non-CANE format scores at most 0.001 across all nine models.

Format design choice. Altering single characters of the format design influences F_1 scores, with the largest gain from the primer (+4.1 F_1). Format group rankings are robust to removing guidelines (Section F.5), varying few-shot count (Section F.3), excluding label normalisation (Section F.4), and changing model version (Section F.6), and are identical across English, German, French, and Chinese on WIKIANN (Section F.7), indicating the findings generalise across languages, while further work is needed to confirm this. Formats that underperform due to compliance issues benefit from reasoning models, but not those with autoregressive generation constraints, or predictions for densely nested documents (Section F.1).

10 Conclusion

Output format is the dominant performance variable in few-shot LLM-based NER, accounting for strict macro F_1 variance approximation of up to 49.1% on simple datasets and 74.9% on complex datasets across 21 formats, 17 datasets (13 simple, 4 complex), and 9 models, exceeding the contribution of model scale and dataset domain. Three distinct failure modes explain F_1 score differences between output format groups, which arise from constraints of autoregressive generation. We introduce the CANE format family, addressing the three properties necessary for robust LLM-based NER. We show that CANE-BASE performs best on flat entities, while CANE variants with discontinuous encoding perform non-trivially on GENIA, whereas all other formats score effectively zero regardless of model type. Together, our findings demonstrate that output format is a primary design variable in generative structured prediction.

Limitations

The evaluation covers nine open-weight models; closed-source models were not evaluated. Because the three failure modes follow from constraints of autoregressive generation rather than from any particular model, and the tier structure is stable across all four model families tested, we expect the format hierarchy to transfer to closed-source and other decoder-based models, though confirming this, along with zero-shot and fine-tuned settings, is left to future work. Whether findings extend to zero-shot or fine-tuned settings is an open question; the two-shot ablation (Section F.3) shows rankings are preserved under reduced example count, and

prior work on fine-tuned NER (Ravikumar et al., 2026; Zhan et al., 2026) reports qualitatively consistent format sensitivity. The cross-lingual evaluation covers four languages on WIKIANN. Rankings are identical across these languages, but the context-anchoring mechanism has not been tested on morphologically rich languages, where tokenization interacts differently with span grounding. Each format has its corresponding parser, and choices about strictness directly affect absolute F_1 values, though not the group-level ordering. Parsers were designed with comparability as the primary criterion; the parser-controlled analysis (Appendix H) isolates the most consequential difference across CANE variants, and absolute scores should be interpreted with this dependency in mind. The variance decomposition uses a group means approximation (Section 6.1) that does not account for format-dataset interactions, which are present and reported in Section 6.3.

Acknowledgments

This work was supported by the Swiss National Science Foundation Project Funding (Grant 10003518). This work was supported as part of the “Swiss AI initiative” by a grant from the Swiss National Supercomputing Center (CSCS) under project ID 175 on Alps. We used Anthropic Claude and GitHub Copilot for writing and coding assistance.

Ethics Statement

All models and datasets used in this work are publicly accessible under their respective licenses or data use agreements. SHARE-2013-P and SHARE-2014-P are derived from clinical text accessed via PhysioNet under a data use agreement (Goldberger et al., 2000); we do not release model outputs or predictions for these datasets. No human subjects research was conducted and no personally identifiable information was collected or processed beyond what is presented in the published corpora.

References

Areej Alhassan, Viktor Schlegel, Monira Aloud, Riza Batista-Navarro, and Goran Nenadic. 2025. [Discontinuous named entities in clinical text: A systematic literature review](#). *Journal of Biomedical Informatics*, 162:104783.

Rina Carines Cabral, Soyeon Caren Han, Areej Alhassan, Riza Batista-Navarro, Goran Nenadic, and Josiah

Poon. 2025. [TriG-NER: Triplet-Grid Framework for Discontinuous Named Entity Recognition](#). *arXiv preprint*. ArXiv:2411.01839 [cs.CL] version: 2.

N. Chinchor and P. Robinson. 1998. [Appendix E: MUC-7 Named Entity Task Definition \(version 3.5\)](#). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. [Introduction to the Bio-entity Recognition Task at JNLPBA](#). In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. [Template-Based Named Entity Recognition Using BART](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Xiang Dai. 2018. [Recognizing complex entity mentions: A review and future directions](#). In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. [Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. [NCBI disease corpus: A resource for disease name recognition and concept normalization](#). *Journal of Biomedical Informatics*, 47:1–10.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. [Context Length Alone Hurts LLM Performance Despite Perfect Retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. [Incorporating Non-local Information into Information Extraction Systems by Gibbs](#)

- Sampling.** In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 193 others. 2025. **Gemma 3 Technical Report.** *arXiv preprint*. Version Number: 1.
- A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. **PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.** *Circulation*, 101(23):E215–220.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. **The Llama 3 Herd of Models.** *arXiv preprint*. Version Number: 3.
- Ralph Grishman and Beth Sundheim. 1996. **Message Understanding Conference- 6: A Brief History.** In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Mohd Ariful Haque, Kishor Datta Gupta, Mohammad Ashiqur Rahman, and Roy George. 2026. **Verbatim Data Transcription Failures in LLM Code Generation: A State-Tracking Stress Test.** *arXiv preprint*. Version Number: 1.
- Han He and Jinho D. Choi. 2023. **Unleashing the True Potential of Sequence-to-Sequence Models for Sequence Tagging and Structure Parsing.** *Transactions of the Association for Computational Linguistics*, 11:582–599.
- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. **OntoNotes: The 90% Solution.** In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.
- Antonín Jarolím, Martin Fajčík, and Lucia Makaiová. 2025. **Can LLMs extract human-like fine-grained evidence for evidence-based fact-checking?** *arXiv preprint*. Version Number: 1.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7B.** *arXiv preprint*. Version Number: 1.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. **Mixtral of Experts.** *arXiv preprint*. Version Number: 1.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. **Cadec: A corpus of adverse drug event annotations.** *Journal of Biomedical Informatics*, 55:73–81.
- Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, Zhiyong Lu, Qingyu Chen, and Hua Xu. 2024. **Advancing entity recognition in biomedicine via instruction tuning of large language models.** *Bioinformatics*, 40(4):btac163.
- Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. **Exploring Nested Named Entity Recognition with Large Language Models: Methods, Challenges, and Insights.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.
- J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. **GENIA corpus—a semantically annotated corpus for bio-textmining.** *Bioinformatics*, 19(suppl_1):i180–i182.
- Jin-Dong Kim. 2026. **Simple Inline PubAnnotation Format.**
- Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun’ichi Tsujii. 2011. **Overview of BioNLP Shared Task 2011.** In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA. Association for Computational Linguistics.
- Jin-Dong Kim and Yue Wang. 2012. **PubAnnotation - a persistent and sharable corpus and annotation repository.** In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205, Montréal, Canada. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. **Efficient Memory Management for Large Language Model Serving with PagedAttention.** In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, Koblenz Germany. ACM.
- Ivan Yee Lee, Loris D’Antoni, and Taylor Berg-Kirkpatrick. 2026. **The Format Tax.** *arXiv preprint*. ArXiv:2604.03616 [cs.CL] version: 1.

- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the Middle: How Language Models Use Long Contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Do Xuan Long, Ngoc-Hai Nguyen, Tivatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. 2025. [LLMs Are Biased Towards Output Formats! Systematically Evaluating and Mitigating Output Format Bias of LLMs](#). 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 299–330, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tengxiao Lv, Ling Luo, Juntao Li, Yanhua Wang, Yuchen Pan, Chao Liu, Yanan Wang, Yan Jiang, Huiyi Lv, Yuanyuan Sun, Jian Wang, and Hongfei Lin. 2025. [A Unified Biomedical Named Entity Recognition Framework with Large Language Models](#). *arXiv preprint*. ArXiv:2510.08902 [cs].
- Danielle Mowery. 2013a. [ShAReCLEF eHealth 2013: Natural Language Processing and Information Retrieval for Clinical Care](#).
- Danielle Mowery. 2013b. [ShAReCLEF eHealth Evaluation Lab 2014 \(Task 2\): Disorder Attributes in Clinical Reports](#).
- Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, and Wendy Chapman. 2014. [Task 2: ShARe/CLEF eHealth Evaluation Lab 2014](#). In *Proceedings of CLEF 2014*, Sheffield, United Kingdom.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. [Cross-lingual Name Tagging and Linking for 282 Languages](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. [Bidirectional LSTM-CRF for Named Entity Recognition](#). In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. [Massively Multilingual Transfer for NER](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.
- Karthik Raman, Iftexhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. [Transforming Sequence Tagging Into A Seq2Seq Task](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11856–11874, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lance Ramshaw and Mitch Marcus. 1995. [Text Chunking using Transformation-Based Learning](#). In *Third Workshop on Very Large Corpora*.
- Lev Ratinov and Dan Roth. 2009. [Design Challenges and Misconceptions in Named Entity Recognition](#). In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.
- Adwait Ratnaparkhi. 1998. [Maximum Entropy Models for Natural Language Ambiguity Resolution](#). PhD thesis, University of Pennsylvania.
- Rishi Ravikumar, Nuhu Ibrahim, and Riza Batista-Navarro. 2026. [Lost in Formatting: How Output Formats Skew LLM Performance on Information Extraction](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5498–5513, Rabat, Morocco. Association for Computational Linguistics.
- Nicole Ren and James Teo. 2025. [SLENDER: Structured Outputs for SLM-based NER in Low-Resource Englishes](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 836–849, Vienna, Austria. 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](#). *arXiv preprint*. ArXiv:2310.03668 [cs.CL].
- Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. [Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment](#). In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting](#). In *International Conference on Learning Representations*. Version Number: 2.

- Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. [Strategies for Span Labeling with Large Language Models](#). *arXiv preprint*. ArXiv:2601.16946 [cs].
- Andrew Shin and Kunitake Kaneko. 2024. [Large Language Models Lack Understanding of Character Composition of Words](#). *arXiv preprint*. ArXiv:2405.11357 [cs.CL].
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. [brat: a Web-based Tool for NLP-Assisted Text Annotation](#). In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.
- Jingtong Su, Jianyu Zhang, Karen Ullrich, Léon Bottou, and Mark Ibrahim. 2025. [A Single Character can Make or Break Your LLM Evals](#). *arXiv preprint*. ArXiv:2510.05152 [cs].
- Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. [Overview of the ShARe/CLEF eHealth Evaluation Lab 2013](#). In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138, pages 212–231. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. [Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models](#). *arXiv preprint*. ArXiv:2408.02442 [cs.CL].
- Mistral AI Team. 2025. [Mistral Small 3](#).
- Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. [WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Simone Tedeschi and Roberto Navigli. 2022. [MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition \(and Disambiguation\)](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Asahi Ushio and Jose Camacho-Collados. 2021. [T-NER: An all-round python library for transformer-based named entity recognition](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.
- 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.
- Yu Wang, Hanghang Tong, Ziyi Zhu, and Yun Li. 2022. [Nested named entity recognition: A survey](#). *ACM Trans. Knowl. Discov. Data*, 16(6).
- Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wieggers, and Zhiyong Lu. 2016. [Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation \(CDR\) task](#). *Database: The Journal of Biological Databases and Curation*, 2016:baw032.
- Faren Yan, Peng Yu, and Xin Chen. 2025. [LTNER: Large Language Model Tagging for Named Entity Recognition with Contextualized Entity Marking](#). In *Pattern Recognition*, pages 399–411, Cham. Springer Nature Switzerland.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *arXiv preprint*. Version Number: 1.
- Qi Zhan, Yile Wang, and Hui Huang. 2026. [Assessment of Generative Named Entity Recognition in the Era of Large Language Models](#). *arXiv preprint*. ArXiv:2601.17898 [cs] version: 1.
- Yidan Zhang and Zhenan He. 2024. [Large Language Models Can Not Perform Well in Understanding and Manipulating Natural Language at Both Character and Word Levels?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11826–11842, Miami, Florida, USA. 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](#). *arXiv preprint*. ArXiv:2308.03279 [cs.CL].

A Entity Structures

Figure 7 illustrates the four entity structures defined in Section 3.

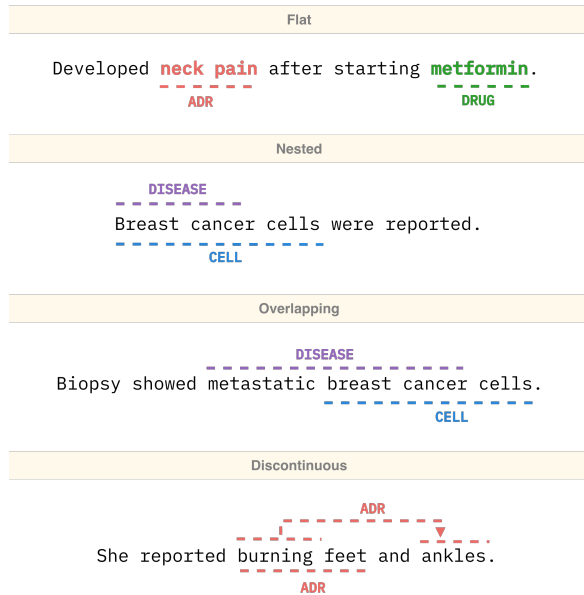


Figure 7: The four entity structures. *Flat*: two contiguous, non-interacting mentions. *Nested*: Breast cancer (DISEASE) is completely contained in Breast cancer cells (CELL). *Overlapping*: metastatic breast cancer (DISEASE) and breast cancer cells (CELL) share tokens, neither containing the other. *Discontinuous*: burning . . . ankles (ADR) is interrupted by a gap.

B Dataset Details

Table 5 lists all datasets used in the evaluation with their canonical paper names and citations. Datasets prefixed with *tner/* were accessed via the T-NER library (Ushio and Camacho-Collados, 2021); all others are identified by their HuggingFace repository name.

Paper name	Dataset identifier	Citation
CoNLL-2003	conll2003	(Tjong Kim Sang and De Meulder, 2003)
NCBI DISEASE	ncbi_disease	(Dogán et al., 2014)
BioNLP-2004	tner/bionlp2004	(Collier et al., 2004)
WNUT-2017	tner/wnut2017	(Derczynski et al., 2017)
BCSCDR	tner/bcscdr	(Wei et al., 2016)
FIN	tner/fin	(Salinas Alvarado et al., 2015)
ONTONOTES-5	tner/ontonotes5	(Hovy et al., 2006)
MULTINERD	tner/multinerd	(Tedeschi and Navigli, 2022)
WIKINEURAL	tner/wikineural	(Tedeschi et al., 2021)
WIKIANN	unimelb-nlp/wikiann	(Rahimi et al., 2019; Pan et al., 2017)
CADEC	cadec	(Karimi et al., 2015)
SHARE-2013-P	share2013	(Mowery, 2013a; Suominen et al., 2013; Goldberger et al., 2000)
SHARE-2014-P	share2014	(Mowery, 2013b; Mowery et al., 2014; Goldberger et al., 2000)
GENIA	genia	(Kim et al., 2003)

Table 5: Dataset names used in the paper, HuggingFace identifiers, and citations. Datasets in the upper block were accessed via T-NER where indicated by the *tner/* prefix.

C Experimental Details

All models were served via vLLM (Kwon et al., 2023) with prefix caching enabled. Standard inference used temperature = 0.7, top-*p* = 0.9, maximum 8,192 model length tokens, and maximum 1,024 output tokens. For reasoning models (Qwen3-1.7B and Qwen3-8B with thinking enabled), temperature was set to 0.6, top-*p* = 0.95, and maximum output tokens to 32,768, following the recommended Qwen3 thinking configuration. Tensor parallelism was set per model: 1 GPU for models up to 7B parameters, 2 GPUs for models between 8B and 30B parameters, and 4 GPUs for LLaMA-3.3-70B and Mixtral-8x7B. Experiments were conducted on a node with 4 NVIDIA GH200 120GB GPUs, totalling approximately 250 GPU hours across all runs. We used a fixed seed of 42.

D Parsing Details

Each model output is parsed by a format-specific parser into a unified internal span representation before evaluation. For *per-token label sequence* formats (Group 1), the parser verifies that the number of generated tags equals the number of whitespace-delimited input tokens; a discrepancy is a critical length mismatch error. For *inline markup* formats (Group 2), spans are extracted from the annotated text copy via stack-based tag or delimiter scanning. For *offset-based* formats (Group 3), character indices are parsed directly from the structured output. For *entity enumeration* formats (Group 4, including SLENDER), spans are resolved by left-to-right string matching against the input. For the CANE family, the context anchor is matched first to identify the intended input window; the entity span is then located within that window, directly operationalising the disambiguation mechanism described in Section 4.2.

E Full Results by Model

Table 6 reports mean strict macro *F*₁ per format per model on simple datasets. Figure 8 shows the *F*₁ distribution per format across all model × dataset combinations; the wide interquartile ranges on CANE variants reflect dataset-level variation rather than instability, as the relative ordering is stable across datasets (Figure 3). Figure 9 shows the rank distribution per format, adding information about spread across experiments.

Format	Gemma-3-4B	LLaMA-3.2-3B	LLaMA-3.3-70B	Mistral-7B-v0.3	Mistral-Small-24B	Mixtral-8x7B	Qwen3-1.7B	Qwen3-30B-A3B	Qwen3-8B	Avg
CANE-BASE	0.442	0.396	0.560	0.264	0.434	0.405	0.346	0.501	0.476	0.425
CANE-FULL	0.413	0.252	0.562	0.385	0.451	0.375	0.314	0.489	0.453	0.411
CANE-CONTEXT	0.409	0.261	0.560	0.376	0.397	0.383	0.306	0.486	0.453	0.404
CANE-DISCO	0.445	0.272	0.559	0.256	0.419	0.384	0.341	0.491	0.455	0.403
SLENDER	0.432	0.316	0.546	0.307	0.363	0.404	0.267	0.451	0.431	0.391
JSON SIMPLE	0.089	0.255	0.544	0.259	0.422	0.312	0.293	0.465	0.435	0.342
INLINE BRACKETED	0.321	0.278	0.500	0.286	0.404	0.286	0.208	0.319	0.451	0.339
GPT-NER	0.349	0.232	0.503	0.275	0.435	0.292	0.201	0.323	0.428	0.338
PUBANNOTATION INLINE	0.341	0.260	0.492	0.252	0.324	0.281	0.190	0.367	0.444	0.328
XML SIMPLE	0.324	0.208	0.480	0.244	0.396	0.254	0.202	0.228	0.409	0.305
XML REF	0.302	0.176	0.436	0.225	0.360	0.241	0.193	0.194	0.374	0.278
STANFORD NER	0.232	0.082	0.387	0.212	0.281	0.279	0.084	0.303	0.288	0.239
STANFORD NER IOB	0.233	0.049	0.419	0.183	0.272	0.287	0.087	0.292	0.288	0.234
SYMBOLIC TAGGING	0.257	0.122	0.462	0.093	0.253	0.183	0.086	0.290	0.320	0.230
IOB	0.187	0.028	0.457	0.079	0.314	0.145	0.091	0.341	0.337	0.220
IOB SENTENCE	0.007	0.007	0.040	0.001	0.017	0.007	0.002	0.025	0.014	0.013
CSV	0.007	0.002	0.041	0.003	0.019	0.007	0.001	0.026	0.012	0.013
JSON	0.000	0.000	0.023	0.005	0.022	0.007	0.003	0.018	0.014	0.010
PUBANNOTATION	0.000	0.001	0.022	0.003	0.021	0.009	0.003	0.019	0.013	0.010
CSV COMPLEX	0.007	0.002	0.020	0.002	0.015	0.006	0.002	0.015	0.008	0.009
BRAT	0.007	0.002	0.018	0.003	0.013	0.008	0.002	0.015	0.007	0.008

Table 6: Mean strict macro F_1 per format per model (simple datasets, 10 corpora). Format ranking is stable across all nine models; the Mistral-7B-v0.3 column shows the anomalous within-family reversal discussed in Section 6.2

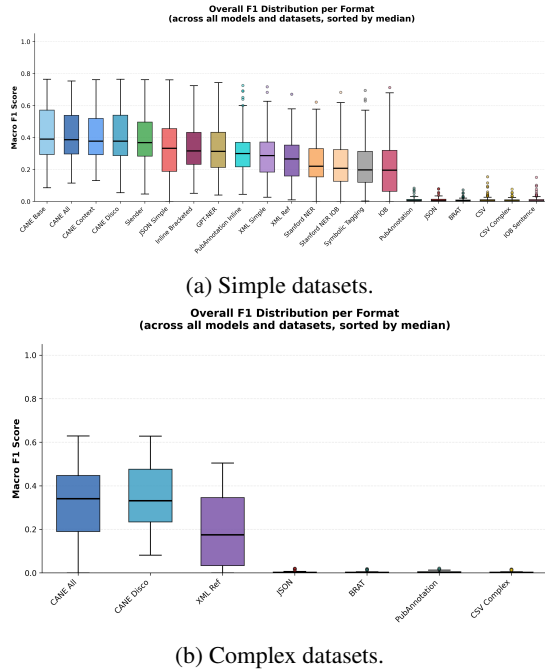


Figure 8: Strict macro F_1 distribution per format across all model $\times$ dataset combinations, sorted by median. Wide interquartile ranges on CANE variants reflect genuine dataset-level variation; the relative ordering is stable across datasets (see Figure 3).

F Supporting Analyses

F.1 Reasoning Models

We compare Qwen3-1.7B and Qwen3-8B with and without reasoning mode across all formats. CANE variants improve at both sizes (1.7B: $+0.083$ – $+0.107$; 8B: $+0.058$ – $+0.079$). At 8B, near-zero formats become mid-tier: JSON rises from 0.013 to 0.217, PUBANNOTATION from 0.013 to 0.220, CSV from 0.010 to 0.198. At 1.7B, near-zero formats remain near-zero. BRAT resists improvement

at both scales ($+0.028$ at 8B), consistent with character counting being a representational rather than a compliance failure. On GENIA, thinking mode reduces CANE-DISCO F_1 at both scales (1.7B: $0.081 \rightarrow 0.056$; 8B: $0.189 \rightarrow 0.157$): longer outputs increase matching attempts, amplifying span errors.

Figures 10 and 11 show per-format F_1 gains from thinking mode on simple and complex datasets, respectively. Table 7 provides the complete per-format results with standard and thinking-mode F_1 side by side.

Format	Qwen3-1.7B (Standard)	Qwen3-1.7B (Thinking)	Δ Qwen3-1.7B	Qwen3-8B (Standard)	Qwen3-8B (Thinking)	Δ Qwen3-8B
Simple Datasets						
CANE-BASE	0.346	0.429	0.083	0.476	0.534	0.058
CANE-DISCO	0.341	0.431	0.091	0.455	0.534	0.079
CANE-FULL	0.314	0.409	0.095	0.453	0.521	0.068
CANE-CONTEXT	0.306	0.414	0.108	0.453	0.530	0.077
JSON SIMPLE	0.293	0.416	0.123	0.435	0.535	0.100
SLENDER	0.267	0.401	0.133	0.431	0.532	0.101
INLINE BRACKETED	0.208	0.201	-0.006	0.451	0.482	0.031
PUBANNOTATION INLINE	0.190	0.225	0.035	0.444	0.493	0.049
GPT-NER	0.201	0.205	0.004	0.428	0.487	0.059
IOB	0.091	0.146	0.055	0.337	0.454	0.118
SYMBOLIC TAGGING	0.086	0.097	0.010	0.320	0.421	0.100
STANFORD NER IOB	0.087	0.130	0.043	0.288	0.374	0.086
STANFORD NER	0.084	0.134	0.051	0.288	0.360	0.073
JSON	0.003	0.004	0.002	0.014	0.217	0.203
IOB SENTENCE	0.002	0.038	0.036	0.014	0.181	0.167
PUBANNOTATION	0.003	0.006	0.003	0.013	0.219	0.206
CSV	0.001	0.006	0.005	0.012	0.200	0.189
CSV COMPLEX	0.002	0.014	0.013	0.008	0.212	0.204
BRAT	0.002	0.004	0.002	0.007	0.035	0.028
Complex Datasets						
CANE-DISCO	0.244	0.274	0.031	0.380	0.453	0.073
CANE-FULL	0.216	0.245	0.029	0.378	0.468	0.090
XML REF	0.104	0.071	-0.033	0.269	0.264	-0.004
PUBANNOTATION	0.001	0.002	0.001	0.004	0.048	0.044
JSON	0.001	0.002	0.001	0.003	0.049	0.046
BRAT	0.001	0.001	0.000	0.002	0.017	0.015
CSV COMPLEX	0.000	0.002	0.001	0.002	0.056	0.054

Table 7: Strict macro F_1 under standard and thinking-mode inference for Qwen3-1.7B and Qwen3-8B per format, averaged across 10 simple datasets. Δ = thinking – standard.

F.2 Discontinuous Demonstration Selection

Replacing one demonstration with four discontinuous examples improves F_1 only for the complex CANE formats: CANE-DISCO gains $+0.006$ ($0.0073 \rightarrow 0.0136$); CANE-FULL gains $+0.003$; XML REF, BRAT, JSON, and PUBANNOTATION show no change. This selection costs 0.008 and 0.021 F_1 on simple and nested entities, favouring

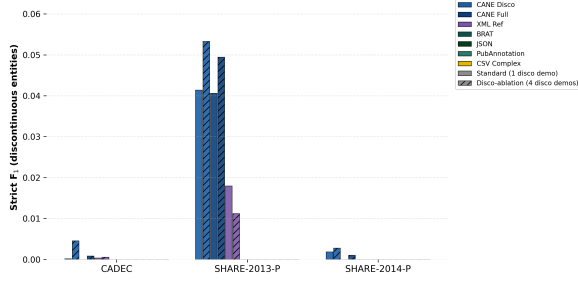


Figure 12: Strict F_1 on discontinuous entities under standard (solid) and disco-ablation (hatched) demonstration selection, per dataset. Only formats with non-zero discontinuous F_1 are visible; offset-based formats score 0.000 in both conditions. SHARE-2013-P shows the largest absolute gains; CADEC and SHARE-2014-P show minimal improvement consistent with lower discontinuous entity density. GENIA is excluded as it contains no discontinuous entities.

do not convey, while natural-language enumeration and inline markup formats benefit (JSON SIMPLE: +0.048; XML REF: +0.034; CANE-DISCO: +0.031), indicating that for formats whose structure is directly exemplified in demonstrations, written guidelines compete with rather than complement examples. The CANE family and SLENDER remain above all other formats in both conditions. All differences are small in absolute magnitude (range -0.011 to $+0.048$), confirming that demonstrations are the primary driver of format compliance and the format specification provides only a marginal additional signal.

F.6 Model Version Sensitivity

Mistral-7B-v0.1 outperforms v0.3 on CANE-BASE (+0.057) and CANE-DISCO (+0.073), while v0.3 outperforms v0.1 on CANE-FULL (-0.111) and CANE-CONTEXT (-0.119), indicating that instruction-tuning changes between versions differentially affected the checkpoint’s response to context-anchored versus non-anchored surface forms. Qwen2.5-7B-Instruct and Qwen3-8B produce comparable results with small absolute differences across most formats (Table 10).

F.7 Cross-Lingual Evaluation

Format rankings are identical across German, French, Chinese, and English on WikiANN. CANE-BASE ranked first and SLENDER ranked fifth in all four languages. Absolute F_1 is lower for Chinese (CANE-BASE: DE=0.594, FR=0.610, ZH=0.413) due to character-level tokenisation interactions with span grounding, but this reduction

Format	5-shot	2-shot	Δ
Simple Datasets			
CANE-BASE	0.425	0.379	-0.046
CANE-FULL	0.411	0.364	-0.046
CANE-CONTEXT	0.404	0.370	-0.033
CANE-DISCO	0.403	0.345	-0.058
SLENDER	0.391	0.362	-0.029
JSON SIMPLE	0.342	0.320	-0.022
INLINE BRACKETED	0.339	0.292	-0.047
GPT-NER	0.338	0.265	-0.073
PUBANNOTATION INLINE	0.328	0.261	-0.067
XML SIMPLE	0.305	0.263	-0.042
XML REF	0.278	0.226	-0.051
STANFORD NER	0.239	0.192	-0.046
STANFORD NER IOB	0.234	0.191	-0.043
SYMBOLIC TAGGING	0.230	0.189	-0.041
IOB	0.220	0.194	-0.026
CSV	0.013	0.013	-0.000
IOB SENTENCE	0.013	0.015	0.001
PUBANNOTATION	0.010	0.010	-0.000
JSON	0.010	0.010	-0.000
CSV COMPLEX	0.009	0.008	-0.001
BRAT	0.008	0.007	-0.001
Complex Datasets			
CANE-DISCO	0.340	0.293	-0.047
CANE-FULL	0.320	0.273	-0.047
XML REF	0.194	0.140	-0.054
PUBANNOTATION	0.004	0.004	-0.000
BRAT	0.003	0.002	-0.000
JSON	0.003	0.003	-0.001
CSV COMPLEX	0.003	0.002	-0.000

Table 8: Mean strict macro F_1 under five-shot and two-shot prompting per format, averaged across 9 models. Δ = two-shot $-$ five-shot.

is uniform across format groups and does not affect rankings.

F.8 Format Family Comparison

Table 12 consolidates the format taxonomy and error analysis, grouping formats by shared structural properties. For each group, we report which of the three properties (P(i)–P(iii)) are satisfied, the dominant failure mode observed empirically, the underlying parser dependency that induces this failure, and the entity structures (flat, nested, overlapping, discontinuous) the format can represent.

G Dispersion of Token-Level Variants

Table 13 lists, for each variant, the mean F_1 over the ten datasets, its standard deviation and range, and how many of the nine models and ten datasets prefer CANE-BASE.

Format	Semantic labels	No semantic labels	Δ
Simple Datasets			
CANE-BASE	0.425	0.418	-0.007
CANE-FULL	0.411	0.407	-0.004
CANE-CONTEXT	0.404	0.400	-0.004
CANE-DISCO	0.403	0.397	-0.005
SLENDER	0.391	0.388	-0.003
JSON SIMPLE	0.342	0.342	0.000
INLINE BRACKETED	0.339	0.336	-0.004
GPT-NER	0.338	0.336	-0.001
PUBANNOTATION INLINE	0.328	0.330	0.002
XML SIMPLE	0.305	0.299	-0.006
XML REF	0.278	0.276	-0.001
STANFORD NER	0.239	0.233	-0.005
STANFORD NER IOB	0.234	0.229	-0.005
SYMBOLIC TAGGING	0.230	0.227	-0.002
IOB	0.220	0.221	0.002
CSV	0.013	0.012	-0.001
IOB SENTENCE	0.013	0.013	0.000
PUBANNOTATION	0.010	0.010	0.000
JSON	0.010	0.011	0.000
CSV COMPLEX	0.009	0.009	-0.000
BRAT	0.008	0.008	-0.000
Complex Datasets			
CANE-DISCO	0.340	0.336	-0.004
CANE-FULL	0.320	0.322	0.003
XML REF	0.194	0.194	0.001
PUBANNOTATION	0.004	0.004	-0.000
BRAT	0.003	0.003	0.000
JSON	0.003	0.003	0.000
CSV COMPLEX	0.003	0.002	-0.000

Table 9: Mean strict macro F_1 with and without semantic label normalisation per format, averaged across 9 models. Δ = no-semantic – semantic.

Dispersion across datasets (0.13 to 0.15) is considerably larger than the differences between variants (0.004 to 0.042), so absolute margins should not be over-interpreted. We therefore assess each effect by its consistency across models and datasets rather than by its margin. Only the primer effect is robust in this sense: it holds across all ten datasets and is largest in magnitude, whereas the separator and delimiter effects are small and model-dependent, with the semicolon separator yielding a higher mean while six of nine models still prefer newlines.

H Parser-Controlled F_1 Comparison

Full-pipeline F_1 measures real-world deployment performance, where parser behaviour is part of the pipeline. The CANE variants differ in how failures propagate (Section 5): CANE-BASE and CANE-DISCO apply document-level failure detection, while CANE-FULL and CANE-CONTEXT apply per-line handling. To separate these two sources of variance, we restrict evaluation to documents where both parsers succeed and recompute F_1 on the intersection subset, providing a view that holds parser strictness constant without superseding the full-pipeline results.

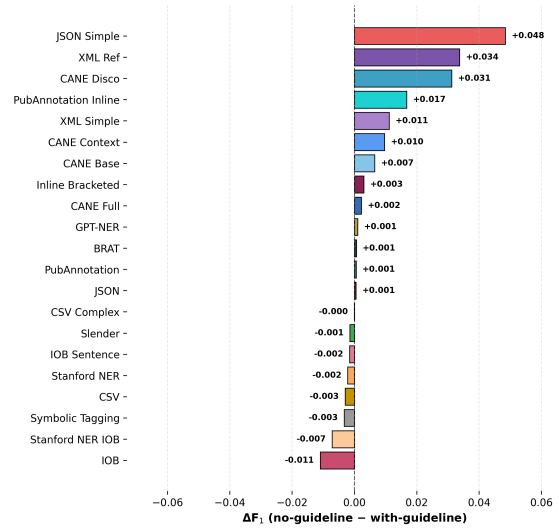


Figure 13: F_1 change from removing the format specification (no-guideline minus with-guideline), averaged across 9 models and 10 simple datasets. Positive values indicate improvement without guidelines; per-token sequence formats are hurt while natural-language enumeration and inline markup formats benefit.

Figure 14 shows the results. The CANE-BASE advantage over CANE-DISCO shrinks from 2.2 to 1.0 points (75% of documents), with the difference attributable to document-level failure propagation in CANE-DISCO. The CANE-BASE advantage over CANE-FULL and CANE-CONTEXT grows from 1.4 to 3.4 and 2.1 to 4.1 points respectively (87% of documents), consistent with per-line handling absorbing failures silently. The full-pipeline ordering CANE-FULL > CANE-DISCO (−0.8 points) reverses on the intersection subset to +2.4 points in favour of CANE-DISCO (78% of documents), replicated on complex datasets (+3.1 points, 77% of documents).

The intersection subset is biased toward structurally simpler documents, so the reported CANE-DISCO advantage is a lower bound on its true advantage across the full distribution.

I Error Examples

Example model outputs illustrating the three failure modes discussed in Section 7: a length mismatch producing an empty prediction, character offset hallucination on a structurally valid output, and a coordinate frame shift caused by capitalization normalization during copying.

Format	Mistral-7B-v0.3	Mistral-7B-v0.1	Δ	Qwen3-8B	Qwen2.5-7B-Instruct	Δ
Simple Datasets						
CANE-FULL	0.385	0.274	-0.111	0.453	0.428	-0.026
CANE-CONTEXT	0.376	0.257	-0.119	0.453	0.430	-0.023
CANE-BASE	0.264	0.321	0.058	0.476	0.465	-0.012
SLENDER	0.307	0.204	-0.103	0.431	0.467	0.036
INLINE BRACKETED	0.286	0.240	-0.046	0.451	0.383	-0.068
CANE-DISCO	0.256	0.329	0.073	0.455	0.469	0.014
GPT-NER	0.275	0.251	-0.024	0.428	0.362	-0.066
PUBANNOTATION INLINE	0.252	0.239	-0.013	0.444	0.376	-0.068
JSON SIMPLE	0.259	0.306	0.047	0.435	0.450	0.015
XML SIMPLE	0.244	0.252	-0.013	0.409	0.329	-0.080
XML REF	0.225	0.206	-0.019	0.374	0.327	-0.047
STANFORD NER	0.212	0.152	-0.060	0.288	0.222	-0.066
STANFORD NER IOB	0.183	0.152	-0.031	0.288	0.239	-0.049
IOB	0.079	0.046	-0.033	0.337	0.278	-0.059
SYMBOLIC TAGGING	0.093	0.096	0.003	0.320	0.272	-0.049
JSON	0.005	0.004	-0.001	0.014	0.009	-0.005
PUBANNOTATION	0.003	0.003	-0.000	0.013	0.009	-0.004
IOB SENTENCE	0.001	0.001	0.000	0.014	0.015	0.001
CSV	0.003	0.001	-0.003	0.012	0.010	-0.002
BRAT	0.003	0.002	-0.001	0.007	0.010	0.002
CSV COMPLEX	0.002	0.001	-0.001	0.008	0.009	0.001
Complex Datasets						
CANE-FULL	0.248	0.132	-0.115	0.378	0.328	-0.050
CANE-DISCO	0.212	0.196	-0.017	0.380	0.364	-0.016
XML REF	0.161	0.088	-0.073	0.289	0.190	-0.098
PUBANNOTATION	0.002	0.000	-0.001	0.004	0.002	-0.002
JSON	0.001	0.001	-0.000	0.003	0.001	-0.002
BRAT	0.000	0.000	0.000	0.002	0.003	0.002
CSV COMPLEX	0.000	0.000	-0.000	0.002	0.002	0.001

Table 10: Mean strict macro F_1 for older and current model checkpoints per format, averaged across 10 simple datasets. Δ = older – current.

Format	EN F_1 Rank	DE F_1 Rank	FR F_1 Rank	ZH F_1 Rank
CANE-BASE	0.558 1	0.594 1	0.610 1	0.413 1
CANE-CONTEXT	0.538 2	0.583 2	0.585 2	0.388 2
CANE-FULL	0.536 3	0.572 3	0.580 3	0.387 3
CANE-DISCO	0.530 4	0.537 5	0.578 4	0.384 4
SLENDER	0.515 5	0.542 4	0.554 5	0.372 5
JSON SIMPLE	0.458 6	0.524 6	0.511 6	0.338 6
GPT-NER	0.445 7	0.445 10	0.412 7	0.309 8
INLINE BRACKETED	0.437 8	0.457 9	0.358 11	0.315 7
PUBANNOTATION INLINE	0.417 9	0.459 8	0.371 9	0.303 10
XML SIMPLE	0.407 10	0.461 7	0.389 8	0.305 9
XML REF	0.375 11	0.422 11	0.359 10	0.302 11
STANFORD NER IOB	0.305 12	0.327 13	0.252 14	0.211 14

Table 11: Mean strict macro F_1 on WikiANN in four languages, averaged across 9 models. Rank in parentheses. Rankings are identical across all four languages. Near-zero (PubAnnotation, BRAT, JSON, CSV Complex) are not listed.

Length mismatch (IOB SENTENCE)

Input (8 tokens): “John Smith flew to Paris on Monday .”
Output (9 tags): “B-PER I-PER O O B-LOC O O O O”
Result: empty prediction list (length mismatch – critical)

Offset hallucination (BRAT)

Input: “Metformin reduces blood glucose in diabetes patients.”
T1 Chemical 0 9 Metformin [correct]
T2 Disease 22 36 blood glucose [hallucinated offsets]
Result: T1 match; T2 invalid_span (critical for T2)

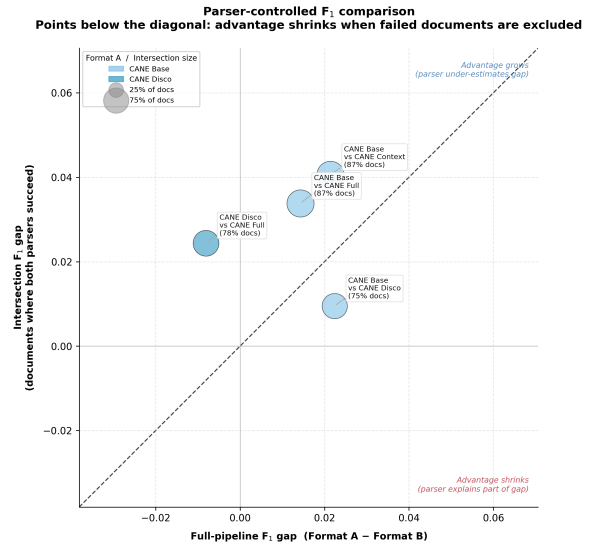


Figure 14: Parser-controlled F_1 comparison for CANE variant pairs (simple datasets, 9 models, 10 datasets). x -axis: full-pipeline F_1 gap (Format A – Format B); y -axis: intersection F_1 gap restricted to documents where both parsers succeed. Bubble area proportional to intersection fraction. Points above the diagonal ($y = x$) indicate the advantage grows when failed documents are excluded; below, it shrinks. The CANE-DISCO vs CANE-FULL bubble has a negative full-pipeline gap but a positive intersection gap, indicating the full-pipeline ordering reflects failure propagation differences rather than format quality.

Coordinate frame shift (GPT-NER)

Original: “...expression of IL-6 in activated T-cells...”
Output: “...expression of @il-6##Protein in activated...”
Result: invalid_span (capitalisation change → offset drift)

J Context Requirements for Entity Disambiguation

Figures 15 and 16 show the proportion of uniquely identifiable entity occurrences as a function of context window size, aggregated across all datasets and per dataset respectively. With no context, 89.5% of occurrences are already unique; a one-token bidirectional window raises this to 98.6% and four tokens reaches 99.9%. The per-dataset breakdown shows that disambiguation difficulty varies substantially. FIN has 52.1% ambiguous occurrences at zero context while MULTINERD and WIKIANN are near-fully unambiguous even without context; the three-token window resolves ambiguity to 99% across all corpora.

Group	Formats	Properties satisfied	Dominant failure mode	Parser dependency	Structures
1. Per-token label sequence	IOB IOB SENTENCE CSV	P(ii)	Length mismatch (0.40–0.94 of predictions)	Token-count alignment: tag count must equal token count	Flat
2. Inline markup	INLINE BRACKETED XML REF GPT-NER	P(i), P(iii)	Verbatim-copy corruption (span error 0.29–0.41)	Regex over generated markup; requires exact input reproduction	Flat Nested Overlapping Discontinuous (XML REF)
3. Offset-based	BRAT PUBANNOTATION JSON CSV COMPLEX	P(iii)	Offset hallucination (span error 1.03–1.36), malformed 0.29	Character-index arithmetic on emitted integers	Flat Nested Overlapping Discontinuous
4. Entity enumeration	JSON SIMPLE PT FORMAT SLENDER	P(i), partial P(ii)	Matching ambiguity on repeated surface forms	Post-hoc string matching	Flat Nested
5. CANE (ours)	CANE-BASE CANE-CONTEXT CANE-DISCO CANE-FULL	P(i), P(ii), P(iii)	Span-matching failures under extreme entity density (GENIA)	String matching; anchored variants add whitespace-flexible context matching	Flat Nested Overlapping Discontinuous

Table 12: Summary of format families, grouping formats by shared properties.

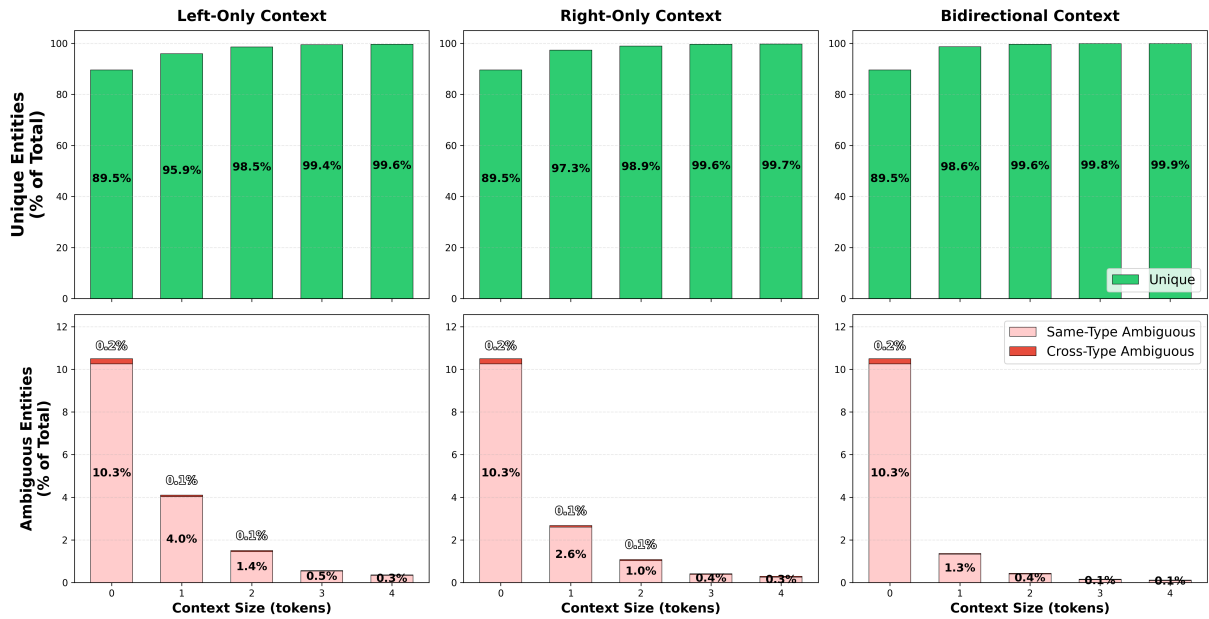


Figure 15: Proportion of uniquely identifiable entity occurrences as a function of context window size, aggregated across all datasets (dataset-equal-weight), for left-only, right-only, and bidirectional context. Upper panel: unique entities as percentage of total. Lower panel: ambiguous entities split by same-type and cross-type ambiguity.

Variant	Mean F1	Std	Min–Max	Base preferred (models, datasets)
CANE Base (newline sep., backticks, primer)	0.424	0.138	0.237–0.593	—
w/ semicolon separator	0.438	0.145	0.229–0.602	6/9, 1/10
w/ double-quote delimiters	0.420	0.142	0.216–0.590	7/9, 6/10
w/o primer header	0.383	0.134	0.195–0.538	7/9, 10/10
PT Format	0.393	0.133	0.192–0.561	6/9, 10/10

Table 13: Dispersion of token-level variants. Std and Min–Max are computed over per-dataset mean F_1 ; *Base preferred* counts how many of the nine models and ten datasets prefer CANE-BASE over the variant.

K Prompt Examples

We provide three representative five-shot prompts. The first shows CANE-BASE on CONLL-2003, illustrating the format on flat entities. The sec-

ond and third show XML REF and CANE-FULL on CADEC, illustrating how a format with and without structural expressivity handles overlapping and discontinuous entities respectively. The prompt structure is identical across all formats and datasets; only the format specification and demonstration outputs vary.



Figure 16: Per-dataset breakdown of entity disambiguation (bidirectional context only). A one-token bidirectional window reduces ambiguous occurrences from 10.5% to 1.4%; four tokens reaches 99.9% unique identification across all 17 datasets.

Five-shot prompt: CANE-BASE on CONLL-2003 (flat entities)

Task: Named Entity Recognition

Entities: LOCATION, MISCELLANEOUS, ORGANIZATION, PERSON

Format: cane_base

Guidelines: Produce a sequence of entity annotations starting with 'Entities:' on the first line, followed by entity annotations where each annotation contains the entity surface text enclosed in backticks, the phrase 'is a' (or 'is an'), and the entity type label. Multiple entity annotations are separated by newlines. This format contains no positional information; entities are identified solely by their surface text and type. When no entities exist in the input, output: NONE

Important: Output only the final answer in the required format. All predictions must exactly match the input verbatim. Do not alter capitalization, do not add or remove whitespace, and do not change punctuation. Do not add explanations, commentary, markdown, or any extra text.

Examples:

Input: .
Output:
NONE

Input: The European Commission said on Thursday it disagreed with German advice to consumers to shun British lamb until scientists determine whether mad cow disease can be transmitted to sheep .
Output:
Entities:
'European Commission' is a ORGANIZATION
'German' is a MISCELLANEOUS
'British' is a MISCELLANEOUS

Input: Fischler proposed EU-wide measures after reports from Britain and France that under laboratory conditions sheep could contract Bovine Spongiform Encephalopathy (BSE) – mad cow disease .
Output:
Entities:
'Fischler' is a PERSON
'EU-wide' is a MISCELLANEOUS
'Britain' is a LOCATION
'France' is a LOCATION
'Bovine Spongiform Encephalopathy' is a MISCELLANEOUS
'BSE' is a MISCELLANEOUS

Input: Germany 's representative to the European Union 's veterinary committee Werner Zwingmann said on Wednesday consumers should buy sheepmeat from countries other than Britain until the scientific advice was clearer .
Output:
Entities:
'Germany' is a LOCATION
'European Union' is a ORGANIZATION
'Werner Zwingmann' is a PERSON
'Britain' is a LOCATION

Input: "What we have to be extremely careful of is how other countries are going to take Germany 's lead , " Welsh National Farmers ' Union (NFU) chairman John Lloyd Jones said on BBC radio .
Output:
Entities:
'Germany' is a LOCATION
'Welsh National Farmers ' Union' is a ORGANIZATION
'NFU' is a ORGANIZATION
'John Lloyd Jones' is a PERSON
'BBC radio' is a ORGANIZATIONv

—
Input: Hartford 4 BOSTON 2
Output:v

Five-shot prompt: XML REF on CADEC (complex entities)

Task: Named Entity Recognition

Entities: ADVERSE_DRUG_REACTION

Format: xml_ref

Guidelines: Reproduce the input text with entity spans enclosed in XML tags containing ref attributes that assign entity identifiers. Each tag uses the entity type as its name and includes a ref attribute with a unique identifier for that entity instance. Discontinuous entities share the same ref value across all their constituent spans. Overlapping entities at identical text positions combine their ref values as space-separated identifiers within a single tag.

Important: Output only the final answer in the required format. All predictions must exactly match the input verbatim. Do not alter capitalization, do not add or remove whitespace, and do not change punctuation. Do not add explanations, commentary, markdown, or any extra text.

Examples:

Input: I have had no side effects been taking Arthrotec a little over a year, have not noticed any side effects.\nIt does help alot I noticed that when there are times when I forget to take it I can't stand or walk for any lengths of time.\n

Output:

I have had no side effects been taking Arthrotec a little over a year, have not noticed any side effects.\nIt does help alot I noticed that when there are times when I forget to take it I can't stand or walk for any lengths of time.\n

Input: I feel a bit drowsy & have a little blurred vision, so far no gastric problems.\nI've been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it.\nDue to my arthritis getting progressively worse, to the point where I am in tears with the agony, gp's started me on 75 twice a day and I have to take it.\nEvery day for the next month to see how I get on, here goes.\nSo far its been very good, pains almost gone, but I feel a bit weird, didn't have that when on 50.\n

Output:

I feel a < ADVERSE_DRUG_REACTION ref="e1"> bit drowsy< ADVERSE_DRUG_REACTION> & have a < ADVERSE_DRUG_REACTION ref="e2"> little blurred vision< ADVERSE_DRUG_REACTION> , so far no < ADVERSE_DRUG_REACTION ref="e3"> gastric problems< ADVERSE_DRUG_REACTION> .\nI've been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it.\nDue to my arthritis getting progressively worse, to the point where I am in tears with the agony, gp's started me on 75 twice a day and I have to take it.\nEvery day for the next month to see how I get on, here goes.\nSo far its been very good, pains almost gone, but I < ADVERSE_DRUG_REACTION ref="e4"> feel a bit weird< ADVERSE_DRUG_REACTION> , didn't have that when on 50.\n

Input: no side effects for the first two months .\nthen vaginal bleeding 2 wks after menstrual cycle.\nstomach pain.\ncanker sores in my mouth.\nheadache.\nbeen off for 1 week still have bleeding.\nhelped my pain alot .\ntoo scared to take this drug again.\n\nOutput:\nno side effects for the first two months .\nthen < ADVERSE_DRUG_REACTION ref="e1"> vaginal bleeding< ADVERSE_DRUG_REACTION> 2 wks after menstrual cycle.\n< ADVERSE_DRUG_REACTION ref="e2"> stomach pain< ADVERSE_DRUG_REACTION> .\n< ADVERSE_DRUG_REACTION ref="e3"> canker sores< ADVERSE_DRUG_REACTION> in my mouth.\n< ADVERSE_DRUG_REACTION ref="e4"> headache< ADVERSE_DRUG_REACTION> .\nbeen off for 1 week still have bleeding.\nhelped my pain alot .\ntoo scared to take this drug again.\n

Input: I can't describe the horrible stomach pain, constipation, uterine, cramps and pain in the ovaries (menopausal for 20 years) pressure in the stomach, couldn't urinate for a whole day, severe stomach pain, constant burping.\nThe experience was a certain kind of hell!\nThis drug is a bad trip!\n

Output:

I can't describe the < ADVERSE_DRUG_REACTION ref="e1"> horrible stomach pain< ADVERSE_DRUG_REACTION> , < ADVERSE_DRUG_REACTION ref="e2"> constipation< ADVERSE_DRUG_REACTION> , < ADVERSE_DRUG_REACTION ref="e3"> uterine, cramps< ADVERSE_DRUG_REACTION> and < ADVERSE_DRUG_REACTION ref="e7"> pain in the ovaries< ADVERSE_DRUG_REACTION> < ADVERSE_DRUG_REACTION ref="e3"> < ADVERSE_DRUG_REACTION ref="e7"> (menopausal for 20 years)< ADVERSE_DRUG_REACTION> < AD-

VERSE_DRUG_REACTION> < ADVERSE_DRUG_REACTION ref="e8"> pressure in the stomach< ADVERSE_DRUG_REACTION> , < ADVERSE_DRUG_REACTION ref="e4"> couldn't urinate< ADVERSE_DRUG_REACTION> for a whole day, < ADVERSE_DRUG_REACTION ref="e5"> severe stomach pain< ADVERSE_DRUG_REACTION> , < ADVERSE_DRUG_REACTION ref="e6"> constant burping< ADVERSE_DRUG_REACTION> .\nThe experience was a certain kind of hell!\nThis drug is a bad trip!\n

Input: It seems as if bleeding forever is a side effect but doctors tell you that its not the drug\nI have been on Arthotec for Three months I still have pain and stiffness\nat first week into this drug I was in so much pain I sat on the floor at work and thought I was giving a baby.\nI always feel weak and almost fainted a bunch of times.\nConsider something that works I am switching today as this isnt in my head.\n

Output:

It seems as if < ADVERSE_DRUG_REACTION ref="e1"> bleeding< ADVERSE_DRUG_REACTION> forever is a side effect but doctors tell you that its not the drug\nI have been on Arthotec for Three months I still have pain and stiffness\nat first week into this drug I was in < ADVERSE_DRUG_REACTION ref="e2"> so much pain< ADVERSE_DRUG_REACTION> I sat on the floor at work and thought I was giving a baby.\nI always < ADVERSE_DRUG_REACTION ref="e3"> feel weak and almost fainted< ADVERSE_DRUG_REACTION> a bunch of times.\nConsider something that works I am switching today as this isnt in my head.\n

Input: Stabbing pain between shoulder blades, I am 52 and feel 82, no energy, muscle tone depletion, strength depletion, numbness in upper legs from knee to thigh, pain and numbness in left foot, left foot larger than right now by one size, pain in feet, need special shoes just to walk now, I feel like I will not be alive next year.\nI've been on this for 5 years, and am going to quit after finding this website.\nDoctors did nothing for me.\nThey have no idea what this drug does to people!\nTime to exercise, eat healthy food, and get OFF this stuff.\nThank you everyone for taking the time to post, and helping me understand what was happening.\nHate this stuff.\nCrestor just as bad.\n

Output:

Five-shot prompt: CANE-FULL on CADEC (complex entities)

Task: Named Entity Recognition

Entities: ADVERSE_DRUG_REACTION

Format: cane_full

Guidelines: Produce output starting with 'Entities:' on the first line, followed by one line per entity as a natural language statement: 'entity_text' is a LABEL found in 'left entity right'. The entity text appears verbatim within the backtick-delimited context string, surrounded by its actual left and right context. For discontinuous entities, list all non-overlapping parts ordered left-to-right with 'and': 'part1' and 'part2' is a LABEL found in 'left1 part1 right1' and 'left2 part2 right2'. Each context string contains the exact text as it appears in the document for that specific part, including the entity part itself. When no entities exist in the input, output: NONE

Important: Output only the final answer in the required format. All predictions must exactly match the input verbatim. Do not alter capitalization, do not add or remove whitespace, and do not change punctuation. Do not add explanations, commentary, markdown, or any extra text.

Examples:

Input: I have had no side effects been taking Arthrotec a little over a year, have not noticed any side effects. It does help alot I noticed that when there are times when I forget to take it I can't stand or walk for any lengths of time.\n

Output:

NONE

Input: I feel a bit drowsy & have a little blurred vision, so far no gastric problems.\nI've been on Arthrotec 50 for over 10 years on and off, only taking it when I needed it.\nDue to my arthritis getting progressively worse, to the point where I am in tears with the agony, gp's started me on 75 twice a day and I have to take it.\nEvery day for the next month to see how I get on, here goes.\nSo far its been very good, pains almost gone, but I feel a bit weird, didn't have that when on 50.\n

Output:

Entities:

'bit drowsy' is a ADVERSE_DRUG_REACTION found in 'a bit drowsy &'

'little blurred vision' is a ADVERSE_DRUG_REACTION found in 'a little blurred vision,'

'gastric problems' is a ADVERSE_DRUG_REACTION found in 'no gastric problems.'

'feel a bit weird' is a ADVERSE_DRUG_REACTION found in 'I feel a bit weird,'

Input: no side effects for the first two months .\nthen vaginal bleeding 2 wks after menstrual cycle.\nstomach pain.\ncanker sores in my mouth.\nheadache.\nbeen off for 1 week still have bleeding.\nhelped my pain alot .\ntoo scared to take this drug again.\n

Output:

Entities:

'vaginal bleeding' is a ADVERSE_DRUG_REACTION found in 'then vaginal bleeding 2'

'stomach pain' is a ADVERSE_DRUG_REACTION found in 'cycle. stomach pain.'

'canker sores' is a ADVERSE_DRUG_REACTION found in 'pain. canker sores in'

'headache' is a ADVERSE_DRUG_REACTION found in 'mouth. headache.'

Input: I can't describe the horrible stomach pain, constipation, uterine, cramps and pain in the ovaries (menopausal for 20 years) pressure in the stomach, couldn't urinate for a whole day, severe stomach pain, constant burping.\nThe experience was a certain kind of hell!\nThis drug is a bad trip!.\n

Output:

Entities:

'horrible stomach pain' is a ADVERSE_DRUG_REACTION found in 'the horrible stomach pain,'

'constipation' is a ADVERSE_DRUG_REACTION found in 'pain, constipation,'

'uterine, cramps' and '(menopausal for 20 years)' are a ADVERSE_DRUG_REACTION found in 'constipation, uterine, cramps and' and 'ovaries (menopausal for 20 years) pressure'

'couldn't urinate' is a ADVERSE_DRUG_REACTION found in 'stomach, couldn't urinate for'

'severe stomach pain' is a ADVERSE_DRUG_REACTION found in 'day, severe stomach pain,'

'constant burping' is a ADVERSE_DRUG_REACTION found in 'pain, constant burping.'

'pain in the ovaries' and '(menopausal for 20 years)' are a ADVERSE_DRUG_REACTION found in 'and pain in the ovaries (menopausal' and 'ovaries (menopausal for 20 years) pressure'

'pressure in the stomach' is a ADVERSE_DRUG_REACTION found in 'years) pressure in the stomach,'

Input: It seems as if bleeding forever is a side effect but doctors tell you that its not the drug\nI have been on Arthotec for Three months I still have pain and stiffness\nat first week into this drug I was in so much pain I sat on the floor at work and thought I was giving a baby.\nI always feel weak and almost fainted a bunch of times.\nConsider something that works I am switching today as this isnt in my head.\n

Output:

Entities:

'bleeding' is a ADVERSE_DRUG_REACTION found in 'if bleeding forever'

'so much pain' is a ADVERSE_DRUG_REACTION found in 'in so much pain I'

'feel weak and almost fainted' is a ADVERSE_DRUG_REACTION found in 'always feel weak and almost fainted a'

Input: Stabbing pain between shoulder blades, I am 52 and feel 82, no energy, muscle tone depletion, strength depletion, numbness in upper legs from knee to thigh, pain and numbness in left foot, left foot larger than right now by one size, pain in feet, need special shoes just to walk now, I feel like I will not be alive next year.\nI've been on this for 5 years, and am going to quit after finding this website.\nDoctors did nothing for me.\nThey have no idea what this drug does to people!\nTime to exercise, eat healthy food, and get OFF this stuff.\nThank you everyone for taking the time to post, and helping me understand what was happening.\nHate this stuff.\nCrestor just as bad.\n

Output:

L Full Results Tables

Tables 15 and 14 report mean strict macro F_1 per format per model on simple and complex datasets respectively, with one column per model and a final column showing the mean across all models. Format rows are sorted by grand average descending.

Format	Gemma-3-4B	LLaMA-3.2-3B	LLaMA-3.3-70B	Mistral-7B-v0.3	Mistral-Small-24B	Mixtral-8x7B	Qwen3-1.7B	Qwen3-30B-A3B	Qwen3-8B	Avg
CADEC										
CANE-DISCO	0.505	0.334	0.584	0.231	0.513	0.467	0.422	0.539	0.522	0.458
CANE-FULL	0.443	0.323	0.593	0.424	0.460	0.426	0.409	0.506	0.517	0.456
XML REF	0.353	0.178	0.437	0.199	0.400	0.289	0.167	0.281	0.343	0.294
PUBANNOTATION	0.000	0.000	0.014	0.003	0.020	0.012	0.003	0.017	0.013	0.009
JSON	0.000	0.000	0.019	0.003	0.020	0.003	0.002	0.017	0.009	0.008
BRAT	0.003	0.001	0.018	0.001	0.015	0.003	0.001	0.014	0.005	0.007
CSV COMPLEX	0.004	0.001	0.016	0.000	0.013	0.005	0.001	0.014	0.005	0.007
GENIA										
CANE-DISCO	0.100	0.088	0.268	0.119	0.238	0.171	0.081	0.234	0.189	0.165
CANE-FULL	0.081	0.038	0.235	0.000	0.154	0.000	0.061	0.191	0.154	0.102
XML REF	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
PUBANNOTATION	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
JSON	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BRAT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CSV COMPLEX	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SHARE-2013-P										
CANE-DISCO	0.329	0.128	0.600	0.249	0.500	0.324	0.218	0.354	0.379	0.342
CANE-FULL	0.308	0.176	0.588	0.247	0.461	0.298	0.185	0.371	0.380	0.335
XML REF	0.198	0.130	0.423	0.172	0.355	0.110	0.120	0.044	0.371	0.214
PUBANNOTATION	0.000	0.000	0.004	0.003	0.009	0.002	0.000	0.005	0.001	0.003
JSON	0.000	0.000	0.006	0.001	0.002	0.003	0.000	0.004	0.002	0.002
BRAT	0.001	0.000	0.004	0.000	0.005	0.003	0.001	0.004	0.001	0.002
CSV COMPLEX	0.001	0.001	0.004	0.000	0.003	0.002	0.000	0.005	0.000	0.002
SHARE-2014-P										
CANE-DISCO	0.395	0.262	0.628	0.249	0.572	0.392	0.253	0.384	0.431	0.396
CANE-FULL	0.379	0.218	0.629	0.320	0.522	0.358	0.208	0.391	0.461	0.387
XML REF	0.244	0.142	0.504	0.274	0.405	0.212	0.128	0.055	0.440	0.267
PUBANNOTATION	0.000	0.000	0.006	0.001	0.007	0.005	0.001	0.009	0.002	0.004
JSON	0.000	0.000	0.006	0.000	0.007	0.003	0.002	0.007	0.002	0.003
BRAT	0.001	0.000	0.006	0.000	0.004	0.002	0.001	0.003	0.000	0.002
CSV COMPLEX	0.001	0.000	0.005	0.001	0.004	0.003	0.000	0.004	0.001	0.002

Table 14: Full results for complex datasets: macro F1 per format and model, grouped by dataset. Each dataset block is preceded by a header row with the dataset name. Formats are ordered by overall average F1 descending. The final column is the mean over all models.

Format	Gemma 3-4B	LLaMA 3.2-3B	LLaMA 3.3-70B	Mistral 7B-v0.3	Mistral Small 24B	Mixtral 8x7B	Qwen3 1.7B	Qwen3 30B-A3B	Qwen3 8B	Avg
BCSCDR										
CANE-BASE	0.573	0.575	0.745	0.417	0.599	0.558	0.485	0.653	0.657	0.585
CANE-FULL	0.484	0.317	0.744	0.519	0.588	0.503	0.448	0.639	0.634	0.542
CANE-CONTEXT	0.457	0.301	0.734	0.521	0.555	0.486	0.437	0.628	0.642	0.529
CANE-DISCO	0.592	0.485	0.744	0.348	0.611	0.525	0.477	0.656	0.649	0.565
SLENDER	0.523	0.520	0.744	0.448	0.484	0.559	0.343	0.600	0.606	0.536
JSON SIMPLE	0.311	0.383	0.761	0.424	0.635	0.362	0.313	0.637	0.593	0.491
INLINE BRACKETED	0.279	0.437	0.725	0.276	0.606	0.349	0.309	0.310	0.616	0.434
GPT-NER	0.433	0.354	0.726	0.234	0.587	0.432	0.290	0.270	0.559	0.432
PUBANNOTATION INLINE	0.331	0.423	0.692	0.195	0.495	0.354	0.281	0.334	0.570	0.408
XML SIMPLE	0.399	0.215	0.718	0.132	0.611	0.340	0.270	0.360	0.493	0.393
XML REF	0.303	0.145	0.573	0.125	0.522	0.319	0.258	0.231	0.373	0.317
STANFORD NER	0.267	0.079	0.536	0.273	0.341	0.489	0.088	0.483	0.445	0.333
STANFORD NER IOB	0.187	0.089	0.550	0.191	0.306	0.425	0.145	0.431	0.392	0.302
SYMBOLIC TAGGING	0.332	0.222	0.694	0.122	0.333	0.288	0.062	0.360	0.503	0.324
IOB	0.289	0.048	0.617	0.156	0.380	0.212	0.173	0.431	0.486	0.310
IOB SENTENCE	0.001	0.010	0.033	0.001	0.010	0.001	0.000	0.041	0.008	0.012
CSV	0.009	0.000	0.036	0.007	0.007	0.001	0.004	0.036	0.007	0.012
JSON	0.000	0.002	0.024	0.006	0.029	0.015	0.004	0.032	0.009	0.013
PUBANNOTATION	0.000	0.000	0.024	0.004	0.027	0.011	0.003	0.025	0.010	0.011
CSV COMPLEX	0.007	0.000	0.021	0.004	0.025	0.008	0.003	0.019	0.010	0.011
BRAT	0.009	0.001	0.018	0.005	0.020	0.004	0.003	0.020	0.008	0.010
BioNLP-2004										
CANE-BASE	0.357	0.246	0.384	0.198	0.349	0.369	0.291	0.354	0.382	0.326
CANE-FULL	0.332	0.116	0.400	0.372	0.348	0.355	0.210	0.362	0.326	0.313
CANE-CONTEXT	0.303	0.131	0.367	0.354	0.325	0.377	0.196	0.337	0.283	0.297
CANE-DISCO	0.350	0.261	0.405	0.226	0.337	0.375	0.252	0.353	0.352	0.324
SLENDER	0.359	0.206	0.316	0.329	0.324	0.398	0.144	0.323	0.311	0.301
JSON SIMPLE	0.007	0.180	0.452	0.188	0.334	0.198	0.190	0.378	0.350	0.253
INLINE BRACKETED	0.300	0.178	0.269	0.334	0.300	0.279	0.129	0.217	0.254	0.251
GPT-NER	0.245	0.095	0.229	0.313	0.296	0.299	0.101	0.211	0.189	0.220
PUBANNOTATION INLINE	0.285	0.181	0.274	0.283	0.284	0.262	0.127	0.317	0.262	0.253
XML SIMPLE	0.235	0.129	0.348	0.271	0.285	0.255	0.115	0.169	0.265	0.230
XML REF	0.281	0.154	0.363	0.269	0.300	0.259	0.129	0.134	0.262	0.239
STANFORD NER	0.162	0.061	0.235	0.129	0.177	0.165	0.003	0.167	0.159	0.140
STANFORD NER IOB	0.162	0.024	0.279	0.124	0.149	0.200	0.003	0.199	0.165	0.145
SYMBOLIC TAGGING	0.123	0.085	0.399	0.094	0.199	0.197	0.040	0.244	0.244	0.181
IOB	0.060	0.002	0.293	0.020	0.154	0.063	0.034	0.198	0.193	0.113
IOB SENTENCE	0.001	0.001	0.002	0.000	0.001	0.000	0.000	0.003	0.000	0.001
CSV	0.004	0.000	0.007	0.000	0.001	0.001	0.000	0.003	0.000	0.002
JSON	0.000	0.000	0.007	0.002	0.004	0.002	0.000	0.005	0.004	0.003
PUBANNOTATION	0.000	0.000	0.005	0.001	0.006	0.001	0.000	0.005	0.002	0.002
CSV COMPLEX	0.000	0.000	0.005	0.000	0.003	0.001	0.000	0.004	0.000	0.001
BRAT	0.000	0.001	0.005	0.000	0.004	0.001	0.000	0.004	0.000	0.002
CoNLL-2003										
CANE-BASE	0.596	0.534	0.765	0.421	0.618	0.594	0.478	0.691	0.614	0.590
CANE-FULL	0.563	0.436	0.754	0.544	0.638	0.545	0.437	0.671	0.593	0.576
CANE-CONTEXT	0.569	0.444	0.762	0.527	0.582	0.557	0.424	0.666	0.582	0.568
CANE-DISCO	0.595	0.402	0.765	0.407	0.565	0.566	0.480	0.672	0.588	0.560
SLENDER	0.567	0.465	0.759	0.483	0.468	0.548	0.426	0.631	0.594	0.549
JSON SIMPLE	0.106	0.393	0.693	0.353	0.555	0.424	0.362	0.564	0.548	0.444
INLINE BRACKETED	0.445	0.344	0.722	0.410	0.463	0.432	0.321	0.544	0.558	0.471
GPT-NER	0.490	0.364	0.744	0.394	0.618	0.378	0.317	0.595	0.538	0.493
PUBANNOTATION INLINE	0.528	0.308	0.690	0.415	0.371	0.425	0.265	0.591	0.544	0.460
XML SIMPLE	0.420	0.240	0.627	0.337	0.424	0.334	0.274	0.333	0.472	0.384
XML REF	0.356	0.178	0.553	0.328	0.368	0.328	0.251	0.316	0.444	0.347
STANFORD NER	0.328	0.186	0.622	0.363	0.401	0.411	0.164	0.434	0.387	0.366
STANFORD NER IOB	0.346	0.110	0.619	0.325	0.389	0.479	0.123	0.467	0.377	0.359
SYMBOLIC TAGGING	0.265	0.182	0.538	0.119	0.312	0.242	0.129	0.304	0.361	0.272
IOB	0.330	0.074	0.680	0.145	0.509	0.300	0.118	0.549	0.462	0.352
IOB SENTENCE	0.010	0.006	0.102	0.005	0.030	0.021	0.003	0.064	0.052	0.033
CSV	0.008	0.006	0.116	0.010	0.035	0.003	0.002	0.048	0.001	0.026
JSON	0.000	0.000	0.080	0.017	0.078	0.021	0.009	0.057	0.051	0.035
PUBANNOTATION	0.000	0.000	0.083	0.014	0.072	0.032	0.010	0.067	0.052	0.037
CSV COMPLEX	0.037	0.004	0.077	0.006	0.057	0.022	0.012	0.052	0.035	0.034
BRAT	0.031	0.007	0.072	0.007	0.054	0.021	0.010	0.050	0.033	0.032
FIN										
CANE-BASE	0.202	0.183	0.315	0.214	0.292	0.226	0.199	0.274	0.255	0.240
CANE-FULL	0.213	0.135	0.366	0.197	0.334	0.229	0.161	0.264	0.271	0.241
CANE-CONTEXT	0.198	0.147	0.363	0.192	0.302	0.205	0.139	0.288	0.291	0.236
CANE-DISCO	0.212	0.084	0.329	0.196	0.249	0.193	0.248	0.266	0.240	0.224
SLENDER	0.188	0.047	0.299	0.118	0.158	0.208	0.122	0.278	0.251	0.185
JSON SIMPLE	0.136	0.145	0.315	0.166	0.203	0.214	0.181	0.201	0.185	0.194
INLINE BRACKETED	0.218	0.052	0.253	0.182	0.282	0.138	0.097	0.230	0.292	0.194
GPT-NER	0.169	0.040	0.306	0.170	0.319	0.174	0.132	0.266	0.288	0.207
PUBANNOTATION INLINE	0.173	0.045	0.262	0.144	0.247	0.188	0.079	0.207	0.265	0.179
XML SIMPLE	0.203	0.027	0.279	0.177	0.265	0.155	0.082	0.154	0.269	0.179
XML REF	0.171	0.011	0.172	0.084	0.217	0.113	0.097	0.124	0.154	0.127
STANFORD NER	0.046	0.000	0.230	0.051	0.219	0.131	0.000	0.202	0.146	0.114
STANFORD NER IOB	0.077	0.000	0.214	0.062	0.198	0.108	0.000	0.140	0.135	0.104
SYMBOLIC TAGGING	0.167	0.053	0.204	0.040	0.218	0.036	0.043	0.100	0.172	0.115
IOB	0.079	0.000	0.269	0.032	0.190	0.047	0.017	0.182	0.210	0.114
IOB SENTENCE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CSV	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
JSON	0.000	0.000	0.007	0.000	0.005	0.000	0.000	0.002	0.001	0.002
PUBANNOTATION	0.000	0.000	0.004	0.000	0.002	0.000	0.002	0.002	0.000	0.001
CSV COMPLEX	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.002	0.000	0.000
BRAT	0.000	0.000	0.003	0.002	0.000	0.000	0.000	0.006	0.000	0.001
MultiNERD										

Format	Gemma 3-4B	LLaMA 3.2-3B	LLaMA 3.3-70B	Mistral 7B-v0.3	Mistral Small 24B	Mixtral 8x7B	Qwen3 1.7B	Qwen3 30B-A3B	Qwen3 8B	Avg
CANE-BASE	0.347	0.356	0.417	0.214	0.286	0.259	0.299	0.395	0.380	0.328
CANE-FULL	0.337	0.212	0.426	0.305	0.287	0.284	0.297	0.408	0.372	0.325
CANE-CONTEXT	0.347	0.200	0.422	0.301	0.187	0.289	0.288	0.393	0.367	0.311
CANE-DISCO	0.352	0.055	0.410	0.227	0.297	0.242	0.300	0.379	0.382	0.294
SLENDER	0.372	0.304	0.424	0.240	0.286	0.274	0.285	0.350	0.340	0.319
JSON SIMPLE	0.000	0.202	0.375	0.159	0.298	0.224	0.263	0.364	0.328	0.246
INLINE BRACKETED	0.339	0.271	0.364	0.332	0.314	0.238	0.186	0.284	0.362	0.299
GPT-NER	0.308	0.209	0.383	0.329	0.330	0.264	0.200	0.314	0.383	0.302
PUBANNOTATION INLINE	0.331	0.235	0.366	0.308	0.295	0.246	0.212	0.330	0.356	0.298
XML SIMPLE	0.322	0.196	0.354	0.288	0.318	0.193	0.245	0.172	0.368	0.273
XML REF	0.324	0.218	0.347	0.265	0.308	0.195	0.225	0.144	0.357	0.265
STANFORD NER	0.165	0.046	0.234	0.164	0.187	0.222	0.091	0.214	0.197	0.169
STANFORD NER IOB	0.220	0.038	0.288	0.157	0.217	0.263	0.106	0.191	0.247	0.192
SYMBOLIC TAGGING	0.270	0.081	0.344	0.075	0.176	0.152	0.121	0.224	0.313	0.195
IOB	0.139	0.003	0.313	0.063	0.236	0.101	0.110	0.241	0.248	0.162
IOB SENTENCE	0.002	0.001	0.013	0.001	0.004	0.001	0.002	0.013	0.005	0.004
CSV	0.011	0.000	0.015	0.000	0.007	0.000	0.001	0.017	0.010	0.007
JSON	0.000	0.000	0.008	0.000	0.003	0.002	0.000	0.003	0.003	0.002
PUBANNOTATION	0.000	0.000	0.005	0.000	0.006	0.004	0.000	0.003	0.003	0.002
CSV COMPLEX	0.002	0.000	0.005	0.000	0.005	0.001	0.000	0.003	0.002	0.002
BRAT	0.001	0.001	0.005	0.000	0.002	0.003	0.000	0.003	0.001	0.002
NCBI Disease										
CANE-BASE	0.411	0.378	0.666	0.293	0.551	0.421	0.306	0.548	0.507	0.453
CANE-FULL	0.398	0.178	0.659	0.336	0.551	0.382	0.294	0.505	0.471	0.419
CANE-CONTEXT	0.378	0.210	0.658	0.315	0.483	0.411	0.298	0.501	0.508	0.418
CANE-DISCO	0.424	0.373	0.659	0.234	0.576	0.400	0.287	0.510	0.442	0.434
SLENDER	0.378	0.303	0.611	0.285	0.456	0.400	0.154	0.459	0.453	0.389
JSON SIMPLE	0.131	0.149	0.644	0.320	0.515	0.457	0.392	0.548	0.508	0.407
INLINE BRACKETED	0.340	0.319	0.645	0.190	0.470	0.232	0.163	0.277	0.517	0.350
GPT-NER	0.267	0.208	0.619	0.206	0.411	0.263	0.185	0.188	0.419	0.307
PUBANNOTATION INLINE	0.345	0.267	0.640	0.151	0.303	0.180	0.132	0.275	0.497	0.310
XML SIMPLE	0.357	0.320	0.626	0.225	0.506	0.277	0.119	0.181	0.475	0.343
XML REF	0.354	0.259	0.559	0.185	0.453	0.237	0.157	0.158	0.433	0.310
STANFORD NER	0.263	0.008	0.464	0.294	0.332	0.313	0.010	0.253	0.302	0.249
STANFORD NER IOB	0.224	0.002	0.520	0.160	0.259	0.226	0.020	0.169	0.273	0.206
SYMBOLIC TAGGING	0.373	0.154	0.629	0.092	0.235	0.271	0.088	0.572	0.516	0.325
IOB	0.094	0.002	0.517	0.037	0.265	0.065	0.025	0.248	0.326	0.176
IOB SENTENCE	0.000	0.006	0.010	0.000	0.004	0.000	0.000	0.004	0.002	0.003
CSV	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.006	0.004	0.002
JSON	0.000	0.000	0.005	0.000	0.007	0.003	0.001	0.005	0.000	0.002
PUBANNOTATION	0.000	0.000	0.005	0.000	0.006	0.005	0.000	0.003	0.002	0.002
CSV COMPLEX	0.000	0.001	0.002	0.001	0.005	0.002	0.000	0.003	0.001	0.002
BRAT	0.000	0.001	0.000	0.001	0.005	0.001	0.001	0.004	0.000	0.001
OntoNotes-5										
CANE-BASE	0.332	0.229	0.475	0.176	0.334	0.293	0.195	0.367	0.339	0.304
CANE-FULL	0.297	0.210	0.444	0.241	0.318	0.260	0.165	0.367	0.317	0.291
CANE-CONTEXT	0.297	0.184	0.457	0.237	0.332	0.265	0.165	0.366	0.331	0.293
CANE-DISCO	0.320	0.222	0.458	0.165	0.323	0.278	0.182	0.362	0.306	0.291
SLENDER	0.319	0.129	0.489	0.215	0.257	0.283	0.117	0.321	0.276	0.267
JSON SIMPLE	0.018	0.157	0.420	0.129	0.285	0.171	0.170	0.305	0.288	0.216
INLINE BRACKETED	0.217	0.157	0.369	0.176	0.284	0.194	0.096	0.243	0.355	0.232
GPT-NER	0.245	0.118	0.375	0.131	0.322	0.149	0.079	0.275	0.321	0.224
PUBANNOTATION INLINE	0.195	0.176	0.360	0.148	0.162	0.168	0.063	0.259	0.349	0.209
XML SIMPLE	0.119	0.061	0.339	0.173	0.306	0.123	0.075	0.120	0.299	0.179
XML REF	0.131	0.055	0.323	0.162	0.280	0.124	0.063	0.112	0.283	0.170
STANFORD NER	0.201	0.046	0.298	0.159	0.224	0.149	0.041	0.199	0.230	0.172
STANFORD NER IOB	0.168	0.016	0.324	0.136	0.211	0.203	0.068	0.195	0.220	0.171
SYMBOLIC TAGGING	0.164	0.028	0.344	0.037	0.206	0.091	0.003	0.191	0.205	0.141
IOB	0.076	0.014	0.315	0.036	0.244	0.063	0.047	0.219	0.250	0.140
IOB SENTENCE	0.001	0.001	0.010	0.000	0.003	0.001	0.000	0.005	0.003	0.002
CSV	0.000	0.001	0.009	0.000	0.008	0.000	0.000	0.005	0.001	0.003
JSON	0.000	0.000	0.025	0.000	0.016	0.004	0.002	0.019	0.011	0.009
PUBANNOTATION	0.000	0.000	0.018	0.001	0.012	0.006	0.001	0.017	0.007	0.007
CSV COMPLEX	0.005	0.001	0.018	0.000	0.018	0.004	0.001	0.021	0.004	0.008
BRAT	0.009	0.001	0.019	0.002	0.009	0.009	0.001	0.018	0.004	0.008
WNUT-2017										
CANE-BASE	0.357	0.278	0.472	0.086	0.343	0.285	0.183	0.404	0.369	0.309
CANE-FULL	0.328	0.201	0.481	0.273	0.362	0.256	0.193	0.410	0.352	0.317
CANE-CONTEXT	0.336	0.221	0.482	0.260	0.325	0.263	0.186	0.401	0.341	0.313
CANE-DISCO	0.361	0.107	0.471	0.100	0.306	0.287	0.176	0.408	0.347	0.285
SLENDER	0.365	0.248	0.445	0.236	0.275	0.300	0.171	0.350	0.294	0.298
JSON SIMPLE	0.172	0.161	0.445	0.145	0.330	0.221	0.197	0.381	0.321	0.264
INLINE BRACKETED	0.233	0.162	0.329	0.191	0.259	0.174	0.119	0.204	0.354	0.225
GPT-NER	0.322	0.190	0.328	0.145	0.254	0.145	0.131	0.216	0.373	0.234
PUBANNOTATION INLINE	0.290	0.168	0.322	0.177	0.228	0.167	0.144	0.256	0.350	0.234
XML SIMPLE	0.285	0.117	0.273	0.157	0.233	0.133	0.161	0.164	0.321	0.205
XML REF	0.255	0.100	0.267	0.172	0.212	0.141	0.131	0.121	0.294	0.188
STANFORD NER	0.136	0.037	0.277	0.142	0.246	0.159	0.029	0.196	0.161	0.154
STANFORD NER IOB	0.131	0.030	0.286	0.122	0.215	0.181	0.072	0.250	0.177	0.163
SYMBOLIC TAGGING	0.137	0.076	0.291	0.052	0.157	0.070	0.031	0.188	0.183	0.132
IOB	0.138	0.015	0.382	0.058	0.240	0.113	0.063	0.333	0.294	0.182
IOB SENTENCE	0.021	0.003	0.041	0.000	0.008	0.005	0.000	0.021	0.013	0.012
CSV	0.008	0.001	0.027	0.000	0.007	0.001	0.000	0.022	0.023	0.010
JSON	0.000	0.000	0.027	0.005	0.017	0.007	0.000	0.014	0.020	0.010
PUBANNOTATION	0.000	0.000	0.027	0.002	0.026	0.008	0.001	0.020	0.012	0.011
CSV COMPLEX	0.001	0.000	0.023	0.001	0.011	0.003	0.000	0.018	0.003	0.007
BRAT	0.000	0.000	0.019	0.000	0.005	0.006	0.000	0.016	0.006	0.006
WikiANN (en)										
CANE-BASE	0.593	0.566	0.608	0.346	0.521	0.581	0.540	0.621	0.647	0.558
CANE-FULL	0.581	0.390	0.611	0.546	0.525	0.512	0.474	0.592	0.596	0.536

Format	Gemma 3-4B	LLaMA 3.2-3B	LLaMA 3.3-70B	Mistral 7B-v0.3	Mistral Small 24B	Mixtral 8x7B	Qwen3 1.7B	Qwen3 30B-A3B	Qwen3 8B	Avg
CANE-CONTEXT	0.593	0.418	0.614	0.537	0.476	0.540	0.472	0.599	0.589	0.538
CANE-DISCO	0.604	0.341	0.609	0.326	0.536	0.553	0.541	0.621	0.639	0.530
SLENDER	0.584	0.500	0.615	0.327	0.464	0.554	0.424	0.575	0.595	0.515
JSON SIMPLE	0.008	0.373	0.611	0.437	0.516	0.476	0.468	0.622	0.614	0.458
INLINE BRACKETED	0.318	0.413	0.602	0.381	0.435	0.458	0.294	0.430	0.598	0.437
GPT-NER	0.452	0.298	0.582	0.380	0.552	0.456	0.270	0.441	0.571	0.445
PUBANNOTATION INLINE	0.363	0.327	0.585	0.361	0.295	0.464	0.255	0.499	0.601	0.417
XML SIMPLE	0.337	0.379	0.553	0.312	0.399	0.420	0.338	0.378	0.552	0.407
XML REF	0.306	0.262	0.534	0.316	0.363	0.396	0.277	0.358	0.561	0.375
STANFORD NER	0.328	0.151	0.398	0.222	0.262	0.336	0.182	0.398	0.387	0.296
STANFORD NER IOB	0.358	0.100	0.427	0.258	0.297	0.363	0.112	0.398	0.429	0.305
SYMBOLIC TAGGING	0.302	0.166	0.535	0.189	0.316	0.223	0.187	0.328	0.278	0.280
IOB	0.340	0.087	0.470	0.163	0.366	0.248	0.095	0.467	0.419	0.295
IOB SENTENCE	0.023	0.041	0.151	0.000	0.095	0.042	0.008	0.077	0.037	0.053
CSV	0.020	0.016	0.155	0.017	0.095	0.059	0.004	0.090	0.048	0.056
JSON	0.000	0.002	0.017	0.014	0.038	0.005	0.007	0.015	0.025	0.014
PUBANNOTATION	0.000	0.004	0.023	0.012	0.035	0.016	0.012	0.018	0.023	0.016
CSV COMPLEX	0.018	0.009	0.021	0.011	0.021	0.008	0.001	0.010	0.022	0.013
BRAT	0.012	0.006	0.021	0.009	0.029	0.026	0.000	0.013	0.017	0.015
WikiNeural										
CANE-BASE	0.652	0.617	0.752	0.271	0.450	0.459	0.483	0.706	0.612	0.556
CANE-FULL	0.601	0.330	0.736	0.518	0.562	0.420	0.461	0.673	0.602	0.545
CANE-CONTEXT	0.613	0.357	0.743	0.513	0.406	0.435	0.457	0.684	0.594	0.533
CANE-DISCO	0.644	0.391	0.735	0.331	0.389	0.421	0.464	0.686	0.608	0.519
SLENDER	0.661	0.443	0.762	0.392	0.453	0.514	0.489	0.626	0.589	0.548
JSON SIMPLE	0.000	0.406	0.725	0.269	0.551	0.371	0.396	0.648	0.616	0.442
INLINE BRACKETED	0.524	0.448	0.722	0.388	0.628	0.364	0.363	0.448	0.606	0.499
GPT-NER	0.561	0.448	0.741	0.448	0.662	0.358	0.305	0.450	0.639	0.513
PUBANNOTATION INLINE	0.565	0.473	0.725	0.339	0.562	0.356	0.341	0.601	0.651	0.513
XML SIMPLE	0.559	0.401	0.683	0.372	0.614	0.311	0.338	0.254	0.575	0.456
XML REF	0.533	0.377	0.671	0.345	0.578	0.301	0.340	0.240	0.580	0.441
STANFORD NER	0.422	0.205	0.578	0.322	0.420	0.417	0.228	0.483	0.460	0.393
STANFORD NER IOB	0.454	0.079	0.682	0.298	0.477	0.421	0.222	0.483	0.462	0.398
SYMBOLIC TAGGING	0.434	0.174	0.640	0.109	0.379	0.262	0.161	0.387	0.427	0.330
IOB	0.322	0.033	0.713	0.079	0.559	0.243	0.230	0.537	0.482	0.355
IOB SENTENCE	0.009	0.000	0.037	0.000	0.011	0.001	0.003	0.028	0.025	0.013
CSV	0.009	0.000	0.039	0.000	0.025	0.001	0.001	0.032	0.024	0.014
JSON	0.000	0.000	0.035	0.002	0.025	0.011	0.002	0.025	0.014	0.013
PUBANNOTATION	0.000	0.001	0.029	0.000	0.018	0.007	0.001	0.028	0.014	0.011
CSV COMPLEX	0.005	0.001	0.028	0.000	0.006	0.006	0.001	0.018	0.006	0.008
BRAT	0.003	0.001	0.023	0.001	0.006	0.004	0.003	0.017	0.005	0.007

Table 15: Full results for simple datasets: macro F1 per format and model, grouped by dataset. Each dataset block is preceded by a header row with the dataset name. Formats are ordered by overall average F1 descending. The final column is the mean over all models.