

Article

Controlled Evaluation of Document-Level Inter-Mention Links for Named-Entity Type Classification in Ukrainian Texts

Oleksii Hryshyn ¹, Volodymyr Shymkovych ^{1,*} and Damian Grela ^{2,*}

¹ Department of Information Systems and Technologies, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Prospect Beresteiskyy 37, 03056 Kyiv, Ukraine; hryshyn.o.s.-it51f@edu.kpi.ua

² Faculty of Electrical and Computer Engineering, Cracow University of Technology, Warszawska 24, 31-155 Cracow, Poland

* Correspondence: v.shymkovych@kpi.ua (V.S.); damian.grela@pk.edu.pl (D.G.)

Abstract

Document-level links between repeated mentions can provide contextual evidence unavailable to a local entity-type classifier, a potentially important capability for Ukrainian because of its limited annotated resources and rich inflectional morphology. However, their specific contribution is difficult to isolate because graph-based systems often simultaneously alter the classifier architecture. We evaluate named-entity type classification with known mention boundaries on NER-UK 2.0 under controlled conditions, using frozen XLM-RoBERTa-base representations, 10 optimization seeds, and 270 graph-head training runs. The evaluation design combines a no-message baseline with matched head parameterization, degree-preserving topology shuffling, simple averaging, oracle diagnostics, and controlled context-weakening interventions; no novel graph architecture is claimed. On intervention-free data, the exact-repeat graph changed macro-F1 by +0.0041 (95% CI [−0.0042, 0.0127]); thus, an overall benefit was not established, although the interval did not rule out a practically meaningful effect. A Ukrainian token-level lemmatization rule increased graph coverage from 34.41% to 42.07% but did not improve performance over exact repetition ($\Delta = -0.0033$, 95% CI [−0.0135, 0.0055]). When only the target mention's context was artificially weakened, with no explicit degradation cue, graph refinement increased target accuracy by 0.0784 (95% CI [0.0544, 0.1021]). The advantage reversed when the entire repeat component was weakened, while simple averaging performed at least as well as the learned graph head. Overall, the results support graph-based refinement as a conditional context-recovery mechanism rather than a universal replacement for local classification. The evidence remains controlled but not independently confirmatory and does not establish gains for end-to-end NER or robustness to naturally degraded context.

Keywords: named-entity recognition; Ukrainian language; graph neural networks; graph-based deep learning; graph-structured data; structured data analysis; document-level graphs; inter-mention relations; message passing; contextual information propagation



Academic Editors: Francisco García-Sánchez and Minoru Sasaki

Received: 1 September 2026

Revised: 6 October 2026

Accepted: 9 October 2026

Published: 13 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Named-entity recognition (NER) is one of the fundamental tasks of natural language processing and an important component of information extraction, enabling automatic detection and semantic classification of entities in unstructured text [1–3]. The principal goal of NER is to locate mentions in text that refer to persons, organizations, geographic objects, dates, events, products, and other domain-specific classes, and assign the corresponding

semantic types to them. Unlike simple keyword search, NER requires contextual analysis to distinguish identically written names and determine their actual semantic referents. NER results are used in relation extraction, semantic search, automatic document summarization, timeline construction, question-answering systems, and the construction of knowledge bases and knowledge graphs: recognized entities can serve as graph nodes, with the relations established between them serving as edges, thereby transforming unstructured text into structured knowledge representations. With the development of machine and deep learning, approaches to NER have evolved from statistical models and BiLSTM-CRF architectures to contextual representations and pretrained transformer language models [4–7]. Recent NER models have also combined pretrained encoders with structured knowledge or syntactic constraints, illustrating that contextual representations and explicitly represented structure can provide complementary signals [8,9]. Nevertheless, the task remains challenging because of multi-token entities, boundary ambiguity, abbreviations, domain-specific terminology, and domain shift. NER therefore serves as an intermediate stage of semantic analysis whose performance directly affects subsequent information extraction and structured-knowledge construction.

For Ukrainian, NER faces additional resource-related and linguistic difficulties arising from morphological variability and the limited availability of high-quality annotated corpora. Inflection of proper names, spelling variation, relatively free word order, and productive abbreviation formation increase the number of entity surface forms, complicating their automatic recognition [10,11]. At the same time, the volume of annotated resources available for training in Ukrainian remains substantially smaller than that available for English, limiting the development of highly accurate NER models in a fully supervised setting [11,12]. Early approaches to Ukrainian NER, in particular, the system of S. Katrenko and P. Adriaans [10], relied largely on manually defined features, dictionary resources, and heuristic rules, which limited their ability to generalize. Subsequent progress in Ukrainian-language resources included the NER-UK 2.0 corpus developed by D. Chaplunskyi and M. Romanyshyn [11], providing a broader foundation for studying modern NER methods. In parallel, the development of multilingual pretrained language models, particularly XLM-RoBERTa, proposed by A. Conneau and co-authors [13], enabled the transfer of contextual language representations to low-resource languages. V. Radchenko and N. Drushchak [12] subsequently evaluated modern language encoders and large language models on NER-UK 2.0. Cross-lingual transfer based on attention and adversarial training provides another example of using supervision from a resource-rich language to support NER in a lower-resource setting [14].

However, even a powerful text encoder does not eliminate all errors when individual mentions are classified primarily from local context without incorporating information from other parts of the document. A transformer encoder produces contextual token representations using a self-attention mechanism that often provides sufficient information for determining the entity type. Difficulties arise for short or contextually incomplete fragments whose immediate surroundings do not contain sufficient lexical or semantic cues for classification. An illustrative example comes from the NER-UK 2.0 corpus, document bruk/f958ce8bb461. In the second sentence of this document, the word “Kyiv” appears in the phrase “(Kyiv, 2015) presents studies by Ukrainian researchers”. Its combination with the year and the surrounding sentence context supports interpretation as a geographic entity, specifically a place of publication. By contrast, in the twelfth sentence of the same document, “Kyiv” appears on a separate line and has no immediate lexical surroundings that could unambiguously determine its type. If such fragments are processed independently, the model has no direct access to the informative contextual cue available elsewhere in the document, even though the same object is referred to in both cases.

This limitation can be mitigated by explicitly establishing links between entity mentions within a document and enabling the exchange of contextual information between them. Such a representation can be formalized as a graph whose nodes are entity mentions and whose edges are predefined or there are learnable links between them. Hereafter, we refer to this representation as the graph representation of the document: unlike merely expanding the textual context, it explicitly specifies the mention pairs permitted to exchange information. A related local-to-global distinction appears in entity-linking research through the combination of local mention context with knowledge-graph structure or document-level entity associations [15,16]. Simply extending the context window solves this problem only partially, since it increases the amount of text the encoder has to process, is limited by its maximum sequence length, and does not explicitly specify which mentions are semantically related. The value of document-level context has been demonstrated in several prior studies. A. Akbik, T. Bergmann, and R. Vollgraf [17] aggregated contextualized representations of identical text strings; S. Schweter and A. Akbik [18] extended the contextual representation to the document level; T. Gui and co-authors [19] used label consistency between mentions within a document; A. Hu and co-authors [20] studied interaction between tokens of multi-word entities; P. Chen and co-authors [21] proposed explicit modeling of links between mentions using graph neural networks.

However, these studies do not unambiguously isolate the contribution of inter-mention links to the observed gains, because the proposed methods often also change the baseline architecture by adding new components and parameters. In particular, in the work of P. Chen and co-authors [21], the comparison contrasts a baseline model with an architecture that additionally contains a graph neural network and a representation-fusion mechanism, so the observed gain reflects not only the use of inter-mention links but also the effect of additional parameters. This leaves open the question of how much of the improvement is attributable to link structure and how much to increased model capacity. For the Ukrainian language, in particular, on the NER-UK 2.0 corpus, the contribution of graph links between repeated entity mentions has not yet been studied systematically.

This paper presents a controlled empirical study of the impact of document structure on named-entity type classification. The graph head, exact matching, neighbor averaging, topology perturbation, and oracle diagnostics are established techniques rather than individually novel components. The contribution is their integrated use to isolate link effects from nominal head parameterization, characterize the mechanism through context interventions, and test exact and morphology-aware link rules for Ukrainian. An overview is presented in Figure 1. To ensure controlled analysis, the problem statement is deliberately narrowed: mention boundaries are considered known, and the module performs only type classification. All experiments use the gold annotated mention spans; in deployment, such spans would have to be supplied by a separate detector. Accordingly, every reported metric is conditional on gold spans and excludes boundary-detection errors. Entity boundary detection, including the detection of nested entities (nested gold spans are retained and classified), coreference resolution, and linking mentions to knowledge-base entries fall outside the scope of this study.

The contributions of this work are empirical and methodological rather than architectural:

1. An integrated controlled design for isolating link effects. A matched-head no-message baseline is combined with degree-preserving topology perturbations, oracle diagnostics, simple averaging, and fixed statistical contrasts. The matched comparison holds nominal parameterization and non-structural inputs fixed while explicitly acknowledging unequal effective prediction paths.

2. A mechanistic test of context transfer. Target-only and whole-component weakening, including matched retraining with the degradation indicator held constant, distinguish access to intact neighboring context from use of an explicit intervention cue.
3. Ukrainian evidence on the coverage–quality trade-off. The study quantifies exact-repeat coverage, audits missed counterparts, and directly compares exact and token-level lemma edge rules on NER-UK 2.0. This evidence concerns type classification with known boundaries, not end-to-end NER.

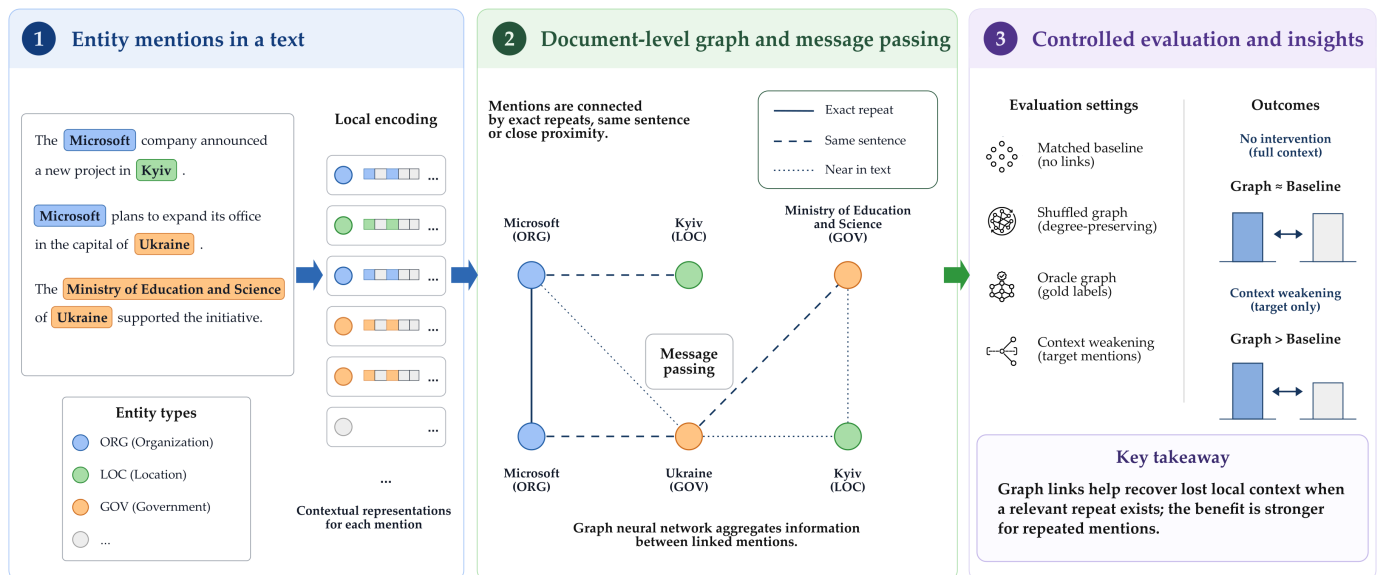


Figure 1. Overall idea of the controlled study. Entity mentions are represented using local contextual information and connected in a document-level graph to enable information exchange between related mentions. Controlled experiments hold the head parameterization, nominal parameter count, and training-data volume fixed and assess the effect of the links on named-entity type classification with known boundaries.

The paper is organized as follows. Section 2 reviews existing approaches to document-level named-entity recognition and the use of inter-mention information, identifies the main methodological limitations of prior studies, and formulates the unresolved problem addressed in the present study. Section 3 defines the aim of the study, specifies the research objectives, and formulates the research questions used to guide the experimental analysis. Section 4 presents the evaluated pipeline for the controlled analysis of mention-link effects and describes its five sequential stages, including mention formation and normalization, local mention encoding, construction of the document graph representation, graph-based type refinement, and model training with result generation. Section 5 describes the experimental study and provides the complete evaluation protocol, including the NER-UK 2.0 corpus and data splits, compared model configurations, controlled graph-topology experiments, training and reproducibility settings, evaluation metrics, statistical analysis, and the controlled context-weakening scenarios. Section 6 presents the experimental results, including the intervention-free evaluation, controlled context-recovery experiments, analysis of training-data volume, distribution of effects across mention groups, and a summary of the answers to the research questions. Section 7 discusses the findings in relation to the research objectives and prior studies, analyzes the role of graph structure and context recovery, identifies the mention groups that benefit most from graph-based refinement, considers the practical implications of the results, and outlines the main limitations of the study. Finally, Section 8 summarizes the principal findings and methodological contri-

butions, states the main conclusions regarding the conditions under which graph-based refinement is effective, and identifies directions for further research.

2. Literature Review and Problem Statement

Classical named-entity recognition systems combined dictionaries, regular-expression rules, morphological features, and statistical sequence models. Their advantage was the interpretability of the features and rules used, but transferring such systems to new domains or text genres often required manual adaptation. Neural models that learn representations directly from text and jointly account for a word's internal structure and its sentence context marked the next stage. In particular, X. Ma and E. Hovy [22] proposed a BiLSTM-CNN-CRF architecture, and G. Lample and co-authors [4] developed a neural architecture for NER combining character-level and contextual representations with BiLSTM-CRF. A systematic review by V. Yadav and S. Bethard [23] showed a gradual transition from manual feature engineering to automatic learning of multi-level representations. Pretrained transformer models constituted the next major stage: BERT, proposed by J. Devlin and co-authors [6], provided bidirectional contextual representations that can be adapted to sequence-labeling and classification tasks. For low-resource languages, multilingual encoders such as XLM-RoBERTa, proposed by A. Conneau and co-authors [13], are particularly important. In this work, the transformer encoder is not a source of novelty; it serves as a shared basis for all compared models, providing identical local representations of mentions so that the experimental analysis can focus on the impact of document links.

For the Ukrainian language, S. Katrenko and P. Adriaans [10] proposed a resource-light approach combining local features and heuristic rules. Subsequent progress in resources for Ukrainian included the NER-UK 2.0 corpus, published by D. Chaplynskyi and M. Romanyshyn [11], which contains named-entity annotation across thirteen classes. An experimental model based on RoBERTa-large, trained on this corpus, achieved strong overall performance—precision 0.90 and recall 0.89—but for rare classes, the results were substantially lower: F1-score was 0.355 for MISC, 0.44 for DOC, and 0.60 for TIME [11]. Subsequently, V. Radchenko and N. Drushchak [12] compared fine-tuned language encoders and large language models on the same corpus. Among the evaluated approaches, the best performance was achieved by a fine-tuned RoBERTa-large-NER model with an F1-score of 0.890, while models based on large language models without dedicated fine-tuning showed lower results: 0.724 for GPT-4o in zero-shot mode and 0.670 for GLiNER [12]. These results indicate that, despite the strong performance of modern pretrained encoders, recognizing rare and contextually ambiguous entities in Ukrainian-language texts remains a challenging problem with clear potential for further improvement.

Document-wide context is incorporated into modern NER systems through several mechanisms that differ in the level and form of interaction between text fragments. The simplest approach is to extend the encoder's context window with information from other parts of the document. S. Schweter and A. Akbik [18], in the FLERT approach, showed that using document features can improve NER quality: for the English CoNLL-03 dataset, F1-score rose from 92.79 to 93.64, and the average gain across five evaluated language datasets was 1.15 percentage points. This approach captures broader context but does not explicitly determine which mentions in the document are relevant to the classification of a particular entity.

Another direction directly exploits repeated mentions. A. Akbik, T. Bergmann, and R. Vollgraf [17] proposed aggregating contextualized representations of identical text strings, so that information from multiple occurrences of the same surface form can inform classification. The proposed approach improved F1-score from 92.86, reported for an earlier model by the same authors [5], to 93.18. The advantage of this approach is the

ability to accumulate contextual information between repeated mentions, but links between them are defined primarily by surface-form identity, which does not cover lexical variation or alternative referring expressions for the same entity.

Another line of research uses document context either to enforce consistency among predicted labels or to refine them. T. Gui and co-authors [19] proposed a mechanism for ensuring label consistency between mentions within a document, which improved F1-score from 91.21 to 92.13. A. Hu and co-authors [20] focused on interactions among tokens within multi-word entities and improved F1-score from 90.92 to 91.92. These works demonstrate that additional structural information can improve entity classification, but they consider different types of dependencies—label consistency or the internal structure of multi-token mentions—and were not designed to isolate the contribution of explicit links between repeated mentions.

Conceptually related approaches model semantic dependencies between mentions or different levels of document context. Z. Dai, H. Fei, and P. Li [24] proposed incorporating coreference links into NER representations, thereby enabling the use of information from mentions that refer to the same entity. Y. Luo, F. Xiao, and H. Zhao [25] developed a hierarchical contextualized representation that accounts for information at different levels, from local context to the document level. Thus, prior research reports benefits from incorporating information beyond the local sentence, but the proposed mechanisms differ substantially in how links between mentions are formed and used. This motivates further study of an explicit document graph with inter-mention links defined independently of the model's other architectural components.

Graph neural networks enable the propagation of information between connected graph nodes and update their representations using both their own features and information from neighboring nodes. A foundational approach is the graph convolutional network proposed by T. Kipf and M. Welling [26]; its aggregation of neighboring information is determined by the graph structure and adjacency matrix. GraphSAGE, proposed by W. Hamilton, R. Ying, and J. Leskovec [27], provides another approach in which the function aggregating information from neighbors is learned directly from data. An important property of GraphSAGE is its inductive formulation, which enables representations to be generated for new nodes and graphs that were not present during training. For document-level NER tasks, this property is essential because each new document induces a separate graph built from its mentions and the links between them. Attention-based graph networks provide a further extension of this approach: P. Veličković and co-authors [28] proposed an architecture that assigns learnable attention coefficients to different neighbors. Graph-based approaches have also been applied to broader information-extraction tasks: Y. Qian and co-authors [29] used a graph structure to model interaction between text elements, while a two-stage mention-level heterogeneous graph attention framework has been proposed for long-distance document-level relation extraction [30].

The work most closely related to the present study is that of P. Chen and co-authors [21], who applied graph neural networks directly to model interactions between entity mentions. The authors formed a graph connecting mentions through coreferential and syntactic links and propagated information between connected nodes using a graph attention network. This approach demonstrated a performance improvement: on the AnatEM corpus, F1-score rose from 83.53 to 84.41, and on the MARS corpus from 89.95 to 90.57 [21]. These results illustrate the potential of explicitly modeling links between mentions. At the same time, the proposed architecture simultaneously introduces graph links and additional neural components, so such a comparison does not isolate how much of the performance gain is attributable specifically to the link structure and how much to increased model capacity. This question is central to the present study.

The main characteristics of the reviewed approaches are summarized in Table 1.

Table 1. Comparison of approaches using links between mentions in named-entity recognition.

Work	Corpus	Mechanism of Using Document Information	Baseline, F1	Model with Links, F1	F1, pts.
(a) Document-level approaches					
Ma, Hovy [22]	CoNLL-2003	sentence-level context, no document links	—	91.21	—
Akbik et al. [17]	CoNLL-03 EN	aggregation of representations of identical text strings	92.86 ± 0.08	93.18 ± 0.09	+0.32
Schweter, Akbik [18]	CoNLL-03 EN	document features and extended context	92.79 ± 0.10	93.64 ± 0.05	+0.85
Gui et al. [19]	CoNLL-2003	label consistency between mentions in a document	91.21	92.13	+0.92
Hu et al. [20]	CoNLL-2003	interaction of tokens of multi-token entities	90.92	91.92	+1.00
Chen et al. [21]	AnatEM	coreferential and syntactic edges, graph attention	83.53 ± 0.32	84.41 ± 0.24	+0.88
Chen et al. [21]	MARS	coreferential and syntactic edges, graph attention	89.95 ± 0.64	90.57 ± 0.47	+0.62
(b) Baseline models for the Ukrainian language					
Chaplynskyi, Romanyshyn [11]	NER-UK 2.0	no explicit document links, RoBERTa-large	—	0.90/0.89 (P/R)	—
Radchenko, Drushchak [12]	NER-UK 2.0	no explicit document links, RoBERTa-large-NER	—	0.890	—

Note: Chaplynskyi and Romanyshyn [11] report precision (P) and recall (R) across all classes, so no directly comparable F1-value is available.

The reviewed works share two common features. First, the performance gain from using document context and links between mentions remains relatively small and, in the experiments reviewed, ranges from 0.32 to 1.00 percentage points of the F1-score. Second, in most studies, the baseline model differs from the model with additional links not only in how contextual information is used but also in architecture, number of layers, or number of parameters. In the present study, we define a matched-parameterization baseline as a model that instantiates the same head and nominal parameters as the corresponding graph model but does not exchange information between mentions. This controls nominal parameterization and non-structural inputs, but not effective predictive capacity: neighbor transformations cannot influence the no-edge model. The comparison is therefore complemented by topology interventions that retain an active neighbor branch while changing or removing its messages. An additional limitation of the reviewed studies is the absence of a graph-structure control based on a random-topology check, with inter-mention links randomly permuted while node degrees are preserved. Such a control experiment determines whether the observed effect depends on meaningful link structure or can be explained merely by propagating and aggregating information between neighboring nodes.

Thus, the unresolved scientific and practical problem is to determine the specific contribution of inter-mention links to mention-type classification performance when boundaries are known, establish the conditions under which such a contribution arises, and identify the groups of mentions for which the use of document structure has a practically significant effect. For the Ukrainian language, in particular, on the NER-UK 2.0 corpus, this problem

has not been systematically studied. Addressing it requires an experimental design that compares models under controlled conditions using a matched-parameterization baseline, a random-topology check of the graph, and controlled weakening of the local context of individual mentions.

3. Aim and Objectives of the Study

The aim of the study is to conduct a controlled evaluation of the effect of inter-mention links on named-entity type classification in Ukrainian-language texts and to establish the conditions under which document-level information transfer is warranted.

Unlike prior studies that typically compare a graph model with a baseline model of a different architecture, the present study holds the head parameterization, nominal parameter count, non-structural inputs, and training-data volume fixed. It combines this matched-parameterization baseline with degree-preserving edge shuffling and controlled weakening of a target mention's local context. Because the neighbor branch is inactive without edges, this design does not claim equality of effective predictive capacity; instead, the topology controls test link placement while the active graph branch is retained.

To achieve this aim, the following objectives were pursued:

- Define a system of metrics and control experiments that evaluates graph-edge effects under fixed nominal head parameterization, non-structural inputs, and training-data volume.
- Implement a conventional graph-refinement head for known mention boundaries within a controlled evaluation that separately analyzes link structure and the availability of local context while explicitly documenting the capacity-control boundary.
- Conduct an experimental study on NER-UK 2.0 using a matched baseline, control topologies, exact and lemma-normalized edge rules, and controlled weakening of local context.
- Establish the conditions under which graph-based refinement is warranted, identify the mention groups showing the strongest effects, and formulate practical recommendations for applying the method.

Each objective corresponds to a research question; the answers are summarized in Section 6.8:

- RQ1: does graph-based refinement improve macro-F1 relative to the baseline with matched head parameterization;
- RQ2: can the graph recover lost local context using information from linked repeated mentions;
- RQ3: how do the edge-construction rules, their class compatibility, and the specific link structure affect the result;
- RQ4: which mention groups exhibit the strongest effect of graph-based refinement.

4. Materials and Methods

The evaluated pipeline consists of five sequential stages: mention formation and normalization, local mention encoding, construction of the document graph representation, type refinement by the graph model, and training and result generation. The overall sequence is given in Table 2.

The task is formalized as follows. Let a document D consist of sentences. After manual annotation, or after processing by a separate boundary-detection module, the set of mentions $\mathcal{M}$ is known. Each mention m_i is specified by a sentence number and start and end positions. Its correct type y_i belongs to a finite set of classes $\mathcal{Y}$. The task is to construct a parameterized function

$$f_{\theta} : (D, \mathcal{M}) \rightarrow (\hat{y}_1, \dots, \hat{y}_n), \quad \hat{y}_i \in \mathcal{Y} \quad (1)$$

that classifies all known mentions. Boundary errors are neither evaluated nor corrected. Stage 1. Mention formation and normalization. This stage converts the document into a representation suitable for graph construction. The inputs are the document text and the known mention boundaries.

Table 2. Overview of the evaluated pipeline for graph-based refinement of named-entity mention types. Mention boundaries are assumed known and come from an external module or annotated corpus; the pipeline performs local encoding, graph construction, and type refinement.

Stage	Steps
Stage 1. Mention formation and normalization	1.1 Sentence segmentation and tokenization; 1.2 Reading mention boundaries; 1.3 Deterministic normalization of the surface form
Stage 2. Local encoding of mentions	2.1 Context-window formation; 2.2 Mention-vector formation; 2.3 Computation of the initial type distribution; 2.4 Node-feature formation
Stage 3. Construction of the document graph representation	3.1 Exact surface-repeat edges; 3.2 Shared-sentence edges; 3.3 Text-proximity edges; 3.4 Rule union and duplicate removal
Stage 4. Type refinement by the graph model	4.1 Projection of node features to a shared dimensionality; 4.2 Computation of the mean neighbor representation; 4.3 Controlled residual update of the node representation; 4.4 Computation of the correction to the local log-probabilities and of the final type distribution
Stage 5. Training and result generation	5.1 Training of the local model; 5.2 Formation of initial local distributions by cross-fitting; 5.3 Separate training of the graph module with weighted cross-entropy loss

Step 1.1. Sentence segmentation and tokenization are performed by a single script using an unchanged configuration for all data splits.

Step 1.2. Reading mention boundaries. The method does not detect boundaries; it receives them from an upstream module or from the annotation.

Step 1.3. Deterministic normalization of the surface form. Normalization is defined as a composition of deterministic operators:

$$\text{norm} = \pi \circ \sigma \circ \lambda \quad (2)$$

where π lowercases the string, σ unifies whitespace and quotation marks, and λ strips outer punctuation. The operator norm is idempotent, i.e., $\text{norm}(\text{norm}(x)) = \text{norm}(x)$, and does not depend on the data split, thereby ensuring identical edge construction for the training, validation, and test splits. Normalization is used only for matching surface forms during graph construction and does not change the textual representation supplied to the encoder. Exact normalization remains the main configuration. A post-review control additionally applies token-level Ukrainian lemmatization, defined below, so that its added coverage and possible analyzer errors can be measured rather than assumed.

This stage yields a set of mentions with normalized forms and preserved boundaries.

Stage 2. Local encoding of mentions. This stage constructs node features that are identical for all compared models. The input is the result of Stage 1.

Step 2.1. Context-window formation. For a mention m_i from sentence s , the window is constructed as

$$w_i = \text{clip}_C(s(m_i), m_i), \quad t_i = \text{trunc}_L(\text{tok}(w_i)) \quad (3)$$

where clip_C limits sentence s to C characters, centering the window relative to the mention if the sentence is longer; tok is subword tokenization, and trunc_L truncates the sequence to L subwords. The resulting sequence is fed to a multilingual transformer encoder with frozen parameters, so the representations h_t are identical for all compared models. The values of C and L and the encoder version are specified in Section 5.4.

Step 2.2. The mention vector is formed by averaging the last hidden states of its subwords and then normalizing by the Euclidean norm:

$$\tilde{c}_i = \frac{1}{|T_i|} \sum_{t \in T_i} h_t, \quad c_i = \frac{\tilde{c}_i}{\max(\|\tilde{c}_i\|_2, \epsilon)} \quad (4)$$

Step 2.3. The initial type distribution is computed by the local classifier:

$$z_i^{loc} = W_2 \text{Dropout}(\text{ReLU}(W_1 c_i + b_1)) + b_2, \quad p_i^{loc} = \text{softmax}(z_i^{loc}) \quad (5)$$

Step 2.4. Node features are formed as follows:

$$x_i = [c_i \parallel p_i^{loc} \parallel r_i \parallel q_i] \quad (6)$$

where r_i consists of three features that do not require graph construction: the number of characters in the mention, the normalized index of its sentence within the document, and a capitalization indicator. The binary indicator q_i marks the known intervention regime: it equals one for artificially weakened mentions in the flagged controlled scenarios and zero in the intervention-free evaluation. For those weakening scenarios, individual models are trained on data in which $q_i = 1$ for weakened training mentions, so this feature is observed during training rather than introduced only at test time. It does not encode the gold label; instead, it informs the model that artificial degradation has been applied. The original controlled experiment therefore evaluates recovery under explicitly flagged weakening rather than robustness to unflagged natural noise.

To quantify dependence on this cue, an otherwise matched constant-indicator control was added. The feature column and full head parameterization were retained, but $q_i = 0$ for every training, validation, and test mention. Both N_0 and G_{rep} were retrained in this condition; the local embeddings, local-classifier probabilities, weakening masks, target selections, optimization seeds, and evaluation rules were identical to those in the flagged condition. Thus, the control removes the explicit cue without introducing a test-only distribution shift or changing the nominal input dimensionality. The number of repeats, node degrees, and other adjacency characteristics are not included in the node features in either condition. The vector p_i^{loc} conveys not only the most likely label but also the shape of the local prediction distribution.

This stage yields the document's node-feature matrix.

Stage 3. Construction of the document graph representation. This stage explicitly defines which mention pairs are permitted to exchange information. For each document, a separate undirected graph $G = (V, E)$ is built, with node $i \in V$ corresponding to mention m_i .

Step 3.1. Exact-repeat edges based on normalized surface form:

$$E_{rep} = \{\{i, j\} : \text{norm}(m_i) = \text{norm}(m_j), i < j\} \quad (7)$$

As an alternative rule, the post-review morphology control tokenizes the normalized mention into word and non-space punctuation tokens. Each word token is replaced by the `normal_form` of the highest-ranked analysis from `pymorphy3` 2.0.6 with the Ukrainian dictionary 2.4.1.1.1663094765; punctuation is retained and the resulting sequence is joined deterministically [31,32]. Denoting this context-free operator by `lem`, the alternative edge set is

$$E_{lem} = \{\{i, j\} : \text{lem}(m_i) = \text{lem}(m_j), i < j\}. \quad (8)$$

Both E_{rep} and E_{lem} form complete undirected cliques within matching groups and are constructed separately in every document without using labels, predictions, or manual audit decisions. The analyzer's first-ranked parse is used without split-specific tuning or correction. Because this analysis is context-free, ambiguous common nouns and proper names can receive incorrect lemmas; the alternative is therefore evaluated as a coverage-quality trade-off rather than treated as ground-truth coreference.

Step 3.2. Shared-sentence edges:

$$E_{sent} = \{\{i, j\} : \text{sent}(m_i) = \text{sent}(m_j), i < j\} \quad (9)$$

Step 3.3. Text-proximity edges are created using a 50-token threshold fixed on the validation sample before test evaluation. The position of a mention is taken as the arithmetic mean of the indices of its first and last tokens, and the distance `dist` as the absolute difference of positions:

$$E_{prox} = \{\{i, j\} : \text{dist}(m_i, m_j) \leq 50, i < j\} \quad (10)$$

Step 3.4. The edge sets are merged and duplicate edges are removed:

$$E_{comb} = E_{sent} \cup E_{rep} \cup E_{prox} \quad (11)$$

If a pair of mentions satisfies several rules, a single edge is kept. No self-loops are added to the graph: the mention's own representation enters the graph layer through a separate self-update and a residual connection (Step 4.3), so the neighbor mean in Step 4.2 contains only other mentions. The graph is built independently for each document, so edges never connect mentions from different documents; this excludes message passing between the training, validation, and test splits and corresponds to the real scenario of processing a new document. The exact-repeat rule only approximates entity identity: identical spelling often, but not always, denotes the same object. Shared-sentence and proximity edges encode textual co-occurrence, not semantic relatedness or coreference, and can connect mentions of different types. None of these rules captures arbitrary paraphrases, aliases, or general semantic relations; the lemma control extends matching mainly to inflectional variants, together with any other forms that receive identical token-level lemmas. The model must therefore learn whether a neighbor's context is useful rather than simply copy its prediction.

This stage yields the document's mention graph.

Stage 4. Type refinement by the graph model. The graph head follows the inductive GraphSAGE principle of neighbor averaging [27] and applies a controlled residual correction to the fixed local distribution. In the final protocol, all compared graphs are passed through a single adjacency channel; the rule type affects the edge composition but is not added as a separate feature.

Step 4.1. Node features are projected to a shared dimensionality:

$$h_i^{(0)} = \text{Dropout}(\text{ReLU}(W_0 x_i + b_0)) \quad (12)$$

Step 4.2. The mean neighbor representation is computed as follows:

$$a_i = \begin{cases} \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} h_j^{(0)}, & |\mathcal{N}(i)| > 0 \\ 0, & |\mathcal{N}(i)| = 0 \end{cases} \quad (13)$$

Step 4.3. The node representation is updated using a controlled residual connection:

$$\begin{aligned} g_i &= \text{sigmoid}(w_g^\top [h_i^{(0)} \| a_i] + b_g), & u_i &= W_s h_i^{(0)} + g_i W_n a_i + b_s \\ h_i^{(1)} &= \text{LayerNorm}(h_i^{(0)} + \text{Dropout}(\text{ReLU}(u_i))) \end{aligned} \quad (14)$$

Step 4.4. The correction to the local log-probabilities, the final type distribution, and the predicted label are computed as follows:

$$\begin{aligned} \gamma_i &= \text{sigmoid}(w_c^\top h_i^{(1)} + b_c), & z_i^{rp} &= s \log \max(p_i^{loc}, \varepsilon) + \gamma_i (W_c h_i^{(1)} + b'_c) \\ p_i^{rp} &= \text{softmax}(z_i^{rp}), & \hat{y}_i &= \arg \max_{y \in \mathcal{Y}} p_{i,y}^{rp} \end{aligned} \quad (15)$$

The scalar g_i controls the contribution of the neighbors within the graph layer, and γ_i controls the magnitude of the final correction. In z_i^{rp} , s is a single learnable scalar that multiplies the vector of local log-probabilities; the logarithm and the maximum with $\varepsilon = 10^{-7}$ are applied element-wise, so zero probabilities do not produce infinite values. The superscript “rp” denotes the refined prediction. The matched-parameterization model N_0 instantiates the same parameters but has no inter-node edges, so $a_i = 0$ for every node. Consequently, W_n and the neighbor gate cannot influence its predictions even though their parameters are included in the nominal total. The contrast with N_0 therefore holds the head parameterization, nominal parameter count, and non-structural inputs fixed, but does not establish equal effective predictive capacity. Degree-preserving topology shuffling and message-removal interventions provide complementary comparisons that retain the trained graph branch.

This stage yields the refined type distribution for each mention.

Stage 5. Training and result generation. The implementation is modular, so the graph module can be integrated into an existing system without redefining mention boundaries.

Step 5.1. The local model is trained on the training split.

Step 5.2. Initial local distributions for the training nodes are generated using a cross-fitting scheme: the training documents are split into several folds, and the prediction for each document is obtained from a model trained without the fold to which that document belongs. Thus, the local model never provides a prediction for a document on which it was trained, so the training features are not optimistically biased. For the validation and test splits, the local model trained on the entire training split is used. The number of folds is given in Section 5.4.

Step 5.3. The graph module is trained separately using a weighted cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{|\mathcal{Y}|} \omega_c \mathbb{I}[y_i = c] \log p_{i,c}^{rp} \quad (16)$$

where N is the number of training mentions; $\mathbb{I}[y_i = c]$ is the correct-label indicator; $p_{i,c}^{rp}$ is the predicted probability of class c ; and ω_c is a class weight inversely proportional to the square root of its frequency in the training split. Training is stopped early based on validation performance.

This stage yields a trained type-refinement model. End-to-end fine-tuning of the encoder is not part of the method, so the observed differences between configurations cannot be attributed to configuration-specific encoder adaptation during training.

5. Experimental Study

5.1. Study Design and Experimental Protocol

The original study comprised two stages. The first stage was an exploratory analysis: on the official test split, the local model and graph variants were compared without matching the number of classifier parameters. Its results were used to generate hypotheses and refine the design. Before the second stage, the set of models, the list of contrasts, the practical-significance thresholds, the seeds, and the analysis rules were fixed; no subsequent changes were made to that second-stage plan. These rules were fixed before the second-stage analysis but were not externally preregistered.

A key limitation is that both original stages used the same official test split. The second stage is therefore a fixed, repeated analysis of a previously inspected test set rather than an independently confirmatory experiment or external replication. This also applies to the practical-significance thresholds: they were fixed before the second stage but selected after the first-stage results had been inspected. The constant-indicator and morphology controls were subsequently designed in response to peer review. For each, the protocol, software versions, seeds, contrasts, and decision rules were fixed before the new models were trained and before the new test outcomes were inspected. Nevertheless, the same test split and the original results were already known, so both additions are controlled post-review analyses rather than independent confirmations. Confidence intervals and *p*-values characterize variability across documents, target-selection realizations, and training runs within this dataset, but they do not account for uncertainty introduced by prior test-set inspection. Independent confirmation requires a new corpus or an untouched split; this limitation is restated in Section 7.5.

We distinguish two concepts below. A model configuration is a combination of classification head architecture, graph regime, and training scenario. A training run is one optimization run of a specific configuration with a specific seed. Absolute metric values are reported as the mean and standard deviation over the training runs of a single configuration, whereas contrasts are reported as differences between the mean metric values of two configurations.

5.2. Data and Preprocessing

The official NER-UK 2.0 corpus in Brat Standoff format [11] was used, fixed at revision 7772b458 07453854 883a8e6d 23e4d145 014a1a42 and distributed under the CC BY-NC-SA 4.0 license. All 13 classes and all nested mentions were retained, so the scheme was not simplified on the basis of prospective test results. In total, 560 documents, 18,339 sentences, and 21,993 mentions were processed. Based on the document-level frequency of each normalized form, 14,960 mentions belong to forms occurring once, 3138 to forms occurring twice, and 3895 to forms occurring at least three times. Splitting was performed only at the document level, so no document or fragment thereof appears in more than one split. The official test split was left unchanged; from the 391 documents of the official split, 333 training and 58 validation documents were formed for development, preserving genre quotas. A five-gram Jaccard-similarity check found no duplicate documents across the splits. The composition of the splits is given in Table 3.

Exact-repeat coverage is the proportion of mentions whose normalized surface form occurs at least twice in the same document; it is not the recall of links between mentions of the same entity. Table 4 reports this coverage by annotated type. The Brat annotations

provide mention spans and types but no entity identifiers or coreference chains, so true same-entity link recall cannot be computed automatically from this corpus.

To characterize aspects not captured by coverage alone, a descriptive manual audit was conducted on 100 uncovered test mentions selected by a fixed SHA-256 ordering before their content was inspected. Complete documents and all same-type mentions were reviewed. No clearly identifiable coreferent was found for 69 mentions; 14 had a clear inflectional variant, 10 an abbreviation or shortened/expanded name, 5 another lexical variant, and 2 were uncertain. Thus, a clear missed coreferent was identified for 29/100 audited mentions (Wilson 95% interval 21.0–38.5%). This single-auditor sample is a descriptive estimate rather than a gold-standard coreference annotation, but it makes the coverage limitation concrete: many uncovered mentions have no identifiable counterpart in the document, while morphology and name variation account for a non-negligible subset of missed links.

To illustrate the experimental application of Stage 3, Figure 2 shows the graph constructed for document bruk/f958ce8bb461. In this corpus example, an exact-repeat edge connects a mention with informative surrounding context to another occurrence lacking lexical context, while shared-sentence edges connect mentions of different types. A substantive discussion of the example and the resulting limitations for Ukrainian is given in Section 7.1.

Document bruk/f958ce8bb461, NER-UK 2.0 corpus

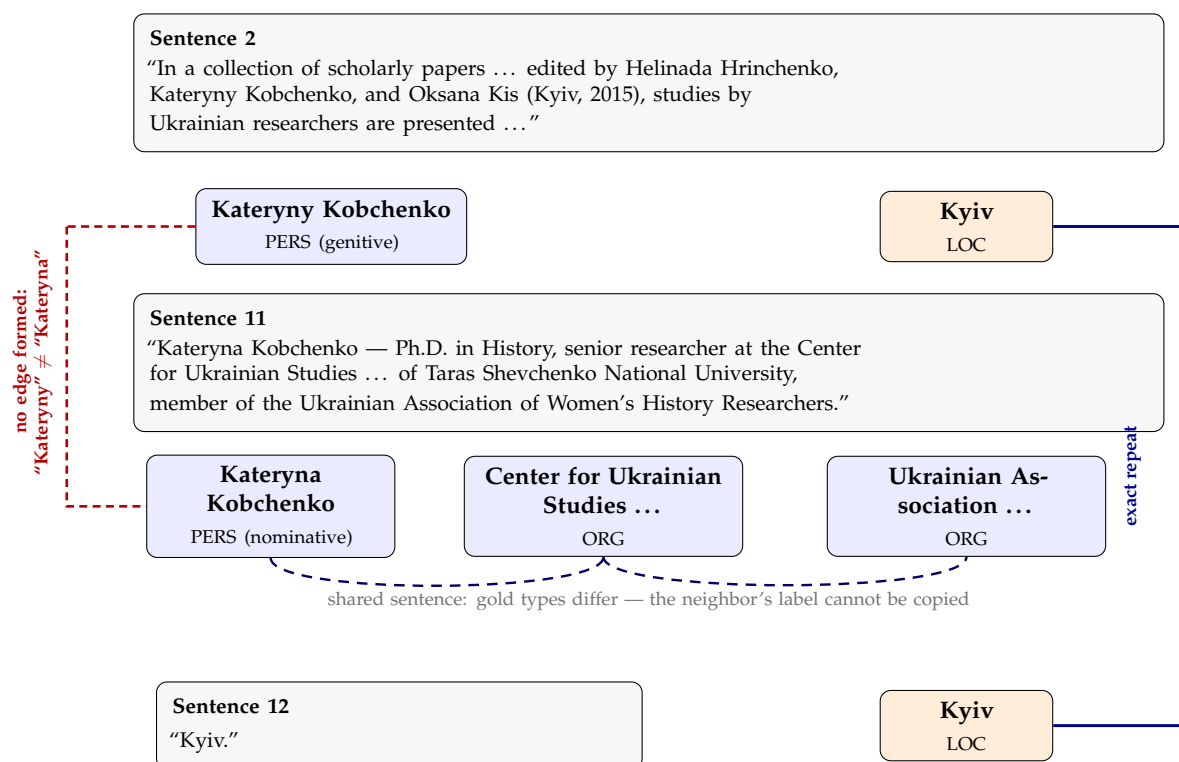


Figure 2. Experimental example of a mention graph constructed using the Stage 3 rules for document bruk/f958ce8bb461 of the NER-UK 2.0 corpus.

Table 3. Composition of the NER-UK 2.0 corpus document splits.

Split	Documents	Sentences	Mentions	Classes
Training	333	10,402	12,799	13
Validation	58	2175	2263	13
Test	169	5762	6931	13

Table 4. Coverage of test mentions by exact normalized-repeat edges, by entity type. Each cell reports covered mentions/all mentions (percentage).

Type	Covered/All	Coverage	Type	Covered/All	Coverage
ART	46/237	19.41%	DATE	53/599	8.85%
DOC	4/40	10.00%	JOB	214/659	32.47%
LOC	254/821	30.94%	MISC	40/142	28.17%
MON	35/325	10.77%	ORG	557/1548	35.98%
PCT	17/90	18.89%	PERIOD	9/185	4.86%
PERS	1138/2186	52.06%	QUANT	18/89	20.22%
TIME	0/10	0.00%	All	2385/6931	34.41%

In the test split, the rare class TIME has 10 mentions, and DOC has 40, so class-level conclusions for them require particular caution.

5.3. Compared Model Configurations

The set of compared configurations and the role of each in the study are listed in Table 5. The division into baseline, trained graph, oracle-diagnostic, and control configurations is part of the evaluation design: without it, the observed gain cannot be attributed to a specific source. The oracle-diagnostic configurations use correct test labels to construct edges or form messages, so they are not suitable for deployment and are used only as gold-informed diagnostic references for the conditional gain under a label-compatible edge composition.

All configurations of the matched family N_0 , G_{rep} , G_{lem} , G_{comb} , G_{clean} , G_{gold} , and G_{shuf} instantiate the same 185,747 nominal parameters, the same single-layer controlled residual head, and the same node features, which do not include graph characteristics. Thus, the contrasts $G_{rep} - N_0$ and $G_{lem} - N_0$ control the nominal parameterization and non-structural inputs. However, they do not equalize effective predictive capacity: with $a_i = 0$, the active neighbor projection and gate cannot affect N_0 predictions. We call the difference of mean metric values between two configurations a contrast. The baseline N_0 was trained separately rather than obtained by removing edges from an already trained graph model; message removal from a trained model is used only as a counterfactual intervention.

Table 6 makes the capacity boundary explicit. The implementation retains four relation branches to keep one shared head definition, but the final single-channel configurations populate only one branch. The three reserved branches are therefore inactive for N_0 , G_{rep} , and G_{lem} . The 128 input-projection weights attached to the q_i column can influence predictions only when q_i varies, that is, in the flagged context-weakening scenarios. Of the 185,747 nominal parameters, 119,183 can therefore influence N_0 predictions and 135,824 can influence G_{rep} predictions in the flagged scenarios, whereas 119,055 and 135,696 can do so whenever q_i is constant: in all intervention-free configurations used for RQ1, E12-M, E14, and E15, and in the constant-indicator control E13-R. In every regime, the graph models have 16,641 additional prediction-influencing parameters. These counts characterize computational paths, not independently estimated statistical degrees of freedom.

The simple alternative A_{rep} does not train a graph head. It post-processes the output distribution $p_i^{N_0}$ of the separately trained N_0 from the same seed and scenario, mixing it with the mean N_0 distribution of the eligible neighbors $\mathcal{N}^*(i) \subseteq \mathcal{N}(i)$:

$$p_i^A = (1 - \alpha)p_i^{N_0} + \alpha \frac{1}{|\mathcal{N}^*(i)|} \sum_{j \in \mathcal{N}^*(i)} p_j^{N_0}, \quad |\mathcal{N}^*(i)| > 0 \quad (17)$$

If $\mathcal{N}^*(i)$ is empty, the N_0 prediction is kept unchanged. On intervention-free data, all nodes are updated and all graph neighbors are eligible; α is selected on validation macro-F1. In the context-weakening scenarios, only target mentions are updated, neighbors whose

context was weakened by the intervention are excluded from $\mathcal{N}^*(i)$, and α is selected on validation target accuracy. This exclusion uses the known intervention mask even when q_i is held constant. It does not affect target-only weakening because each target's repeat neighbors are intact by construction. Under whole-component weakening, however, every target loses all eligible neighbors, so A_{rep} coincides with N_0 by definition rather than as an empirical result. The coefficient α is chosen from the fixed grid $\{0, 0.25, 0.50, 0.75, 1\}$ separately for each graph regime, scenario, and seed. A_{lem} applies the same intervention-free procedure to E_{lem} . The graph G_{clean} is obtained by removing from G_{comb} all edges between mentions with different gold types. In G_{gold} , mentions are ordered within each document–gold-type pair and connected to the mentions one and two positions later in this order, so the maximum undirected degree equals four. Gold-label voting replaces the local distribution with the frequency distribution of the neighbors' gold labels, and in the absence of neighbors, leaves the N_0 prediction unchanged. All three constructions use test labels, i.e., they encode part of the answer in the edges or messages, and serve only as oracle diagnostics that are unavailable at deployment. They evaluate the conditional potential of the model given label-compatible links, but do not show that the corresponding links can be recovered from features available without gold labels.

Table 5. Compared models and their role in the study.

Notation	Role	Edges	Parameters	What It Tests
M_{loc}	lower reference bound	none	other head	quality of a simple head over the frozen encoder
N_0	matched baseline model	none	185,747	quality of the matched head without message passing
A_{rep}	simple averaging of N_0 outputs	exact repeats	no added parameters	whether a complex network is needed to transfer context
A_{lem}	simple averaging control	lemma repeats	no added parameters	effect of morphology-aware coverage without a learned graph head
G_{rep}	trained graph model	exact repeats	185,747	contribution of links between exact repeats
G_{lem}	morphology control	lemma repeats	185,747	effect of a morphology-aware alternative edge rule
G_{comb}	broader graph	repeats, sentence, proximity	185,747	impact of a broader but noisier edge composition
G_{clean}	oracle diagnostic (not deployable)	union after gold-label cleaning	185,747	joint effect of filtering and altered graph structure
G_{gold}	oracle diagnostic (not deployable)	sparse edges by gold labels	185,747	potential under label-compatible links
G_{shuf}	random-topology control	shuffled, degree-preserving	185,747	whether the specific links matter

Graph structure is controlled using degree-preserving randomization: edges are shuffled only within a document so that the node set V , the number of edges $|E|$, and the degree of every node are preserved, while the semantic correspondence between neighboring mentions is disrupted; self-loops and repeated edges are not allowed. Five predefined topology realizations with seeds 907, 1009, 1103, 1201, and 1301 were used. Edge label homogeneity is defined as

$$\text{hom}(E) = \frac{|\{(i, j) \in E : y_i = y_j\}|}{|E|} \quad (18)$$

Because of the high initial homogeneity of the exact-repeat graph, even the shuffled graph remains fairly homogeneous. This control is therefore conservative and should be interpreted as a degree-preserving random-topology control rather than as a fully label-neutral random graph.

Table 6. Layer-wise parameter breakdown for the matched-parameterization head. “Yes” means that the group can influence predictions in the specified configuration.

Parameter Group	Parameters	N_0	G_{rep}/G_{lem}
Input projection W_0, b_0 , excluding the q_i column	100,480	yes	yes
Input-projection weights on the q_i column	128	if q_i varies	if q_i varies
Self update W_s, b_s	16,512	yes	yes
Active neighbor projection W_n	16,384	no	yes
Active neighbor gate w_g, b_g	257	no	yes
Three reserved inactive relation branches	49,923	no	no
Layer normalization	256	yes	yes
Correction projection W_c, b'_c	1677	yes	yes
Correction gate w_c, b_c	129	yes	yes
Local-logit scale s	1	yes	yes
Nominal total	185,747	185,747	185,747
Prediction-influencing total, q_i varies (flagged weakening)	—	119,183	135,824
Prediction-influencing total, q_i constant (all other regimes)	—	119,055	135,696

5.4. Training Setup and Reproducibility

A frozen FacebookAI/xlm-roberta-base model was used as the encoder; the encoding batch size was 48 mentions. Freezing was a deliberate control choice: every downstream configuration received identical local representations, so changes in the measured contrasts could not be attributed to configuration-specific encoder adaptation. The local context window was limited to $C = 900$ characters and the tokenized sequence to $L = 256$ subwords (Step 2.1). To generate the initial local distributions of the training nodes (Step 5.2), the training set was split into five document folds. The local multilayer perceptron used a hidden dimensionality of 128, dropout of 0.15, a learning rate of 0.005, a weight-decay coefficient of 10^{-4} , and was trained for 120 epochs. The graph head used one controlled residual layer, hidden dimensionality 128, layer normalization (LayerNorm), dropout of 0.20, a learning rate of 0.003, a weight-decay coefficient of 10^{-4} , a maximum of 200 epochs, early stopping after 25 epochs without improvement, and gradient-norm clipping at 5. Adam, W [33] was used for both heads.

Ten optimization seeds were used (17, 29, 43, 59, 71, 89, 101, 113, 127, 149). In the original second-stage protocol, 20 graph-head configurations were trained per seed: four training-data-volume scenarios (10%, 25%, 50%, 100%) and two controlled-weakening scenarios (surface-context weakening and replacement with the prior distribution)—six training scenarios in total; in each scenario N_0 , G_{rep} , and G_{gold} were trained (3 models $\times$ 6 scenarios = 18 configurations), and in the full-data-volume scenario G_{comb} and G_{clean} were additionally trained (2 more configurations). The constant-indicator analysis added four heads per seed: N_0 and G_{rep} retrained separately with the indicator enabled and held constant. Local models and out-of-fold probabilities were shared between the two indicator regimes within a seed. The morphology control added three intervention-free heads per seed— N_0 , G_{rep} , and G_{lem} —trained together with $q_i = 0$ and validation macro-F1 as the early-stopping criterion. The revised total is therefore 270 graph-head training runs:

$$200 + 10 \times (2 \text{ indicator regimes}) \times (2 \text{ heads}) + 10 \times (3 \text{ morphology-control heads}) = 270. \quad (19)$$

Five realizations of target-mention selection and controlled corruption (101, 211, 307, 401, 503) and five realizations of random topology (907, 1009, 1103, 1201, 1301) were applied when evaluating already-trained models and did not involve retraining. Simple averaging A_{rep} and A_{lem} add no trainable parameters to the separately trained N_0 ; G_{shuf} uses the trained G_{rep} model with a shuffled topology. The averaging coefficient was selected per seed from $\{0, 0.25, 0.50, 0.75, 1\}$ on validation macro-F1 in intervention-free scenarios and on validation target accuracy in context-weakening scenarios. Computation was performed on an NVIDIA GeForce RTX 4070 SUPER GPU. The post-review runs used Python 3.12.14, NumPy 2.3.5, PyTorch 2.13.0 with CUDA 13.0, and Transformers 5.14.1; the morphology run additionally fixed pymorphy3 2.0.6 and pymorphy3-dicts-uk 2.4.1.1.1663094765. Configuration and source checksums, checkpoints, selections, predictions, and structural statistics were stored with the run manifests.

Run settings, the mention lists for each group, the target-mention selection, predictions, model states, and raw metric values were all stored. Reproducibility checks included the document split, feature identity, the absence of self-loops and repeated edges after shuffling, and the invariance of the test split.

5.5. Metrics and Statistical Analysis

We use the following metric definitions: precision—the fraction of predictions assigned to a class that are correct; recall—the fraction of true instances of a class that are correctly identified; accuracy—the fraction of mentions whose type is classified correctly; macro-F1—the macro-averaged F1-score across classes. To avoid conflating concepts, macro-averaged precision and recall are hereafter called macro-precision and macro-recall.

The main metric for the intervention-free test is macro-F1. This choice reflects the fact that the corpus contains classes with substantially different frequencies (from 40 mentions of class TIME to 6235 mentions of class PERS), and macro-averaging gives equal weight to every class regardless of its frequency; this is particularly important since the rare classes have the greatest room for improvement. Accuracy, macro-precision, macro-recall, and per-class metrics are additionally considered. For the controlled tasks, the main metric is accuracy on 800 predefined target mentions.

The practical-significance thresholds are 0.010 for macro-F1 and 0.005 for accuracy. They were fixed before the original second-stage analysis, after the first-stage test results had been inspected; the same accuracy threshold was fixed in the written constant-indicator protocol before the new control models were trained. These thresholds were not externally registered. They represent the minimum point-estimate changes considered sufficient to justify additional graph-based processing: 0.010 corresponds to the upper end of the 0.0032–0.0100 gains reported for document-level NER in Table 1, and 0.005 means at least half a percentage point of accuracy. We report separately whether a 95% confidence interval excludes zero and whether the point estimate exceeds the relevant threshold; neither criterion is described as independent confirmation.

The contrast of two configurations is computed as

$$\Delta(A, B) = F(A) - F(B) \quad (20)$$

where $F(\cdot)$ is the value of the corresponding metric. The point estimate is the mean paired difference over optimization seeds. Uncertainty was estimated with a paired crossed bootstrap: in replicate b , optimization seeds and test documents are independently resampled with replacement, documents are stratified by text source, and corruption realizations are resampled within a seed; the same bootstrap samples are used for all components of a

given contrast. The bootstrap realization of the contrast and the 95% confidence interval are defined as

$$\Delta^{(b)}(A, B) = F^{(b)}(A) - F^{(b)}(B), \quad b = 1, \dots, B, \quad B = 20,000 \quad (21)$$

$$\text{CI}_{95\%} = [Q_{0.025}(\Delta^{(b)}), Q_{0.975}(\Delta^{(b)})] \quad (22)$$

where Q_q is the q -th sample quantile of the bootstrap distribution. For macro-F1, in each replicate, the per-document confusion matrices were summed and the metric was re-computed over all 13 classes; mentions were not resampled independently. Two-sided p -values were computed from the centered null distribution with a correction for the finite number of replicates. The Holm correction [34] was applied separately within the five original second-stage families E11–E15. The constant-indicator protocol defined a separate three-contrast primary family: the unflagged $G_{rep} - N_0$ gain, the indicator-dependence interaction, and the target-only versus whole-component interaction. The morphology protocol defined a separate two-contrast primary family: $G_{lem} - G_{rep}$ and $G_{lem} - N_0$ in macro-F1. Its 0.010 practical threshold, ten seeds, 20,000 bootstrap replicates, analyzer versions, and full-test evaluation mask were fixed before the lemma graph was constructed or new models were trained. Corrected values are denoted p_{adj} . Reference contrasts outside these families are identified as secondary or descriptive. For each reported secondary p_{adj} value, the Holm correction was applied within the corresponding predefined secondary family, as specified in the text. Absolute estimates are reported without selecting the best run [35]. All metric values and contrasts are reported to four decimal places, and p -values to three.

5.6. Controlled Experiments with Context Weakening

The intervention-free test assesses overall practical performance but provides only a weak test of context transfer: most mentions are already correctly classified locally, and exact-repeat edges cover only 34.41% of test mentions. Controlled tasks were therefore added. In each of the 800 exact-repeat groups, one target mention was selected without using the label; the selection rule is based on the SHA-256 hash function, with five predefined realizations. The quantitative characteristics of these groups, which determine potential scope for graph-based information transfer, are given in Table 7.

Table 7. Characteristics of the controlled task on the 800 target mentions of the test split.

Characteristic	Value
Number of exact-repeat groups in the test split	800
Number of target mentions (one per group)	800
Number of mentions in repeat components	2385
Fraction of test mentions covered by repeat edges	34.41%
Exact-repeat edges in the test split	5008
Label homogeneity of the exact-repeat graph	0.9970
Groups with disagreement among gold repeat types	12 (1.50%)
Fraction of target mentions misclassified by N_0 with clean context	9.67%
Fraction of target mentions misclassified by N_0 after context weakening	17.80%

The last two rows characterize the potential for recovery: after weakening the context of the target mention, the matched baseline model errs on approximately 17.8% of targets, and in almost all groups (98.50%) at least one undamaged repeat has the same gold type and can therefore potentially provide useful context.

For each weakening type, configurations N_0 , G_{rep} , and G_{gold} were trained separately: in the training and validation splits, one target mention from each repeat group was weakened

and q_i was set to 1, so the intervention indicator was observed during training. In the target-context-weakening regime, the text of the mention itself is preserved, but the external context in its window is replaced with blanks, while other repeats remain clean. In the component-context-weakening regime, the same model trained under target weakening is evaluated without retraining, but context is removed and $q_i = 1$ is set for all mentions of the component. The difference between the regimes tests the dependence of the target mention's prediction on undamaged evidence in its neighbors under explicitly flagged intervention. It does not evaluate robustness to unflagged natural context degradation, and the component-weakening pattern is used only as a diagnostic test because it changes the one-weakened-target training pattern to multiple weakened neighbors. In the strictest regime (prior substitution), the vector representation of the target mention is zeroed and its local probability vector is replaced by the class-frequency distribution computed exclusively from the training data.

The constant-indicator control repeats the target-context and component-context conditions with $q_i = 0$ for every mention during training, validation, and testing; the prior-substitution regime was not repeated. N_0 and G_{rep} are retrained separately in both the flagged and constant-indicator regimes using the same 10 optimization seeds. The five test target-selection realizations, training target seed 314,159, validation target seed 161,803, weakening operations, early-stopping rule, and 800-target evaluation mask are unchanged. A_{rep} is retained as a reference without additional trained parameters; α is selected on validation target accuracy separately for each indicator regime and seed. As described in Section 5.3, it excludes weakened neighbors using the intervention mask; this matters only for whole-component weakening, causing A_{rep} to equal N_0 by construction. The flagged heads in this run also provide a same-environment replication of the original E13 contrast. Because the official test split had already been inspected, the control is interpreted as a post-review controlled robustness analysis.

The protocol covers the five original experiment groups E11–E15 and the two post-review controls E12-M and E13-R, each testing a specific stage of the method (Table 8).

Table 8. Correspondence between experiment groups and method stages.

Group	What It Tests	Composition	Method Stage
E11	oracle potential of label-compatible links	gold-label graph, averaging over the gold graph, gold-label voting	Stage 3—gold-informed diagnostic reference
E12	practical quality and the effect of edge composition	N_0 , repeat graph, rule union, cleaned graph, averaging	Stage 3, Steps 3.1–3.4
E12-M	morphology-aware alternative edge rule	matched N_0 , exact-repeat and lemma-repeat graphs, separate averaging controls	Stage 1 normalization and Stage 3.1
E13	mechanism of context restoration	clean context, target weakening, component weakening, prior-distribution substitution, message removal; degree-preserving shuffling reported descriptively	Stage 4—message passing
E13-R	dependence on the degradation indicator	matched q_i enabled versus held-constant retraining	Stage 2 feature control and Stage 4 message passing
E14	dependence on training-data volume	nested sequence of subsamples 10%, 25%, 50%, 100% preserving class fractions	Stage 5
E15	groups in which the effect is concentrated	repeated and singleton, uncertain and confident, their intersection, training-ambiguous forms	Stages 3–4

For E15, a mention is considered repeated if its normalized form occurs at least twice in the same document; the remaining mentions are singletons. Uncertainty is defined by the entropy of the mean N_0 distribution over the ten seeds,

$$H_i = - \sum_{y \in \mathcal{Y}} \bar{p}_{i,y} \log \bar{p}_{i,y} \quad (23)$$

The uncertainty threshold equals the 75th percentile of H_i on the validation set and, once fixed, is applied to the test set without change; the complement of this group forms the confident-mentions group. Training-ambiguous normalized forms are those with support of at least 5 and an entropy of the distribution of gold types of at least 0.5 bits, computed only on the training split. The predefined interactions “repeated vs. singleton” and “uncertain vs. confident” belong to family E15; the intersection of uncertain and repeated mentions and the group of training-ambiguous forms are reported descriptively, without separate inferential claims.

6. Study Results

This section presents the original fixed second-stage results and the constant-indicator and morphology controls added during review. The main comparisons use the matched-parameterization baseline N_0 , which instantiates the same nominal head parameters and receives the same non-structural inputs as the graph model but has no information exchange between mentions. As Table 6 shows, this does not imply equal effective capacity. Results on intervention-free test data are presented first, including the alternative lemma edge rule, followed by diagnostic oracle analyses, controlled context-weakening experiments, topology and training-volume controls, and the distribution of effects across mention groups. The observed patterns are interpreted in Section 7.

6.1. Overall Quality on the Intervention-Free Test

Results on the intervention-free test are given in Table 9.

Table 9. Macro-F1 on the intervention-free test, mean $\pm$ standard deviation over 10 training runs (group E12).

Configuration	Role	Macro-F1	vs. N_0
M_{loc}	lower reference bound	0.7143 ± 0.0219	−0.1036
N_0	matched baseline model	0.8179 ± 0.0124	—
A_{rep}	simple averaging	0.8210 ± 0.0127	+0.0031
G_{rep}	trained graph model	0.8220 ± 0.0144	+0.0041
G_{comb}	broader graph	0.8119 ± 0.0109	−0.0060
G_{clean}	oracle diagnostic—not deployable	0.8430 ± 0.0055	+0.0251

Note: The lower reference bound M_{loc} is a simple classification head of a different architecture over the frozen encoder; it shows how much quality depends on the head itself and is not a matched baseline model. The row G_{clean} uses correct test labels to filter edges and is reported only as a diagnostic estimate.

For the deployable exact-repeat graph, the contrast $G_{rep} - N_0$ is +0.0041 in macro-F1. Its 95% confidence interval $[-0.0042; 0.0127]$ crosses zero, $p_{adj} = 0.651$, and the point estimate is smaller than the 0.010 practical-significance threshold fixed before the second-stage analysis. Thus, the intervention-free test does not establish an overall graph benefit. It also does not establish the absence of a practically meaningful effect, because the interval extends above 0.010. The broader graph G_{comb} likewise has no established gain: $G_{comb} - N_0 = -0.0060$, with interval $[-0.0146; 0.0028]$ and $p_{adj} = 0.446$.

A shift between macro-precision and macro-recall was observed at the exploratory stage, when the compared models did not have matched parameter counts. For the local model of that stage, macro-precision was 0.8378, macro-recall 0.7019, macro-F1 0.7261;

for a two-layer graph model with full edge union, the corresponding values were 0.8149, 0.8391, and 0.8253. This pattern suggests that the more complex classifier assigned rare classes more frequently, identifying more correct mentions at the cost of more false-positive predictions. These values were obtained under a different configuration (three seeds, a two-layer graph head without parameter matching) and are not comparable to the values in Table 9; because of the small number of test mentions in classes DOC (40) and TIME (10), the corresponding changes remain exploratory.

6.2. Morphology-Aware Alternative Edge Rule

The post-review E12-M control compares exact normalized repetition with the token-level lemma rule defined in Section 4. Table 10 shows that lemmatization adds 2767 undirected edges and connects 531 mentions that are isolated under exact matching, raising test coverage by 7.66 percentage points. The largest absolute increases occur for PERS (223 newly covered mentions), JOB (101), LOC (86), and ORG (84). Gold-type homogeneity remains high, although it is a descriptive diagnostic rather than an input to graph construction or an estimate of coreference precision.

Table 10. Structural effect of the alternative lemma edge rule on the test split (group E12-M). Coverage denotes participation in at least one within-document edge; homogeneity is computed from gold types only after graph construction.

Edge Set	Undirected Edges	Covered Mentions	Coverage	Label Homogeneity
Exact normalized repeat, E_{rep}	5008	2385	34.41%	0.9970
Lemma-normalized repeat, E_{lem}	7775	2916	42.07%	0.9959
Added by lemmatization	2767	531 newly covered	+7.66 points	0.9939

All heads in E12-M were trained together using the same local representations, optimization seeds, validation rule, and $q_i = 0$ throughout. Consequently, the within-control comparisons in Table 11 do not mix checkpoints from the original E12 run, and its absolute values are not directly comparable with Table 9. E12-M evaluates the alternative edge rule only on intervention-free data; context-weakening experiments were not repeated with lemma edges. Despite its higher coverage, G_{lem} does not establish an improvement over G_{rep} : the primary paired macro-F1 contrast is -0.0033 , with 95% interval $[-0.0135; 0.0055]$ and $p_{adj} = 0.954$. The second primary contrast, $G_{lem} - N_0$, is $+0.0030$, interval $[-0.0053; 0.0117]$, $p_{adj} = 0.954$. Thus, neither a morphology-aware advantage over exact repetition nor a benefit over the matched no-edge head is established.

Table 11. Intervention-free performance in the matched morphology control, mean $\pm$ standard deviation over 10 optimization seeds (group E12-M).

Configuration	Macro-F1	Accuracy	Macro-F1 vs. N_0
N_0	0.8153 ± 0.0110	0.9126 ± 0.0038	—
A_{rep}	0.8178 ± 0.0113	0.9163 ± 0.0042	+0.0026
A_{lem}	0.8188 ± 0.0110	0.9175 ± 0.0039	+0.0036
G_{rep}	0.8215 ± 0.0152	0.9170 ± 0.0057	+0.0063
G_{lem}	0.8182 ± 0.0122	0.9167 ± 0.0050	+0.0030

As a secondary reference, $A_{lem} - N_0 = +0.0036$ macro-F1 (95% interval $[0.0009; 0.0067]$, $p_{adj} = 0.040$ after Holm correction within the predefined two-contrast secondary macro-F1 family $\{A_{lem} - N_0, G_{lem} - A_{lem}\}$). The interval excludes zero but remains entirely below the 0.010 practical threshold. The evidence therefore supports a structural conclusion—deterministic lemmatization expands coverage while retaining high class

compatibility—but not a practically meaningful predictive improvement on this reused test split.

6.3. Diagnostic Evaluation of the Potential of Structural Information

To estimate the potential performance gain from inter-mention links, graphs built from gold labels were constructed (Table 12). These configurations use correct test labels that are not available in a real system and provide a diagnostic estimate of the potential value of structural information, rather than an estimate of the achievable quality of a practical system.

Table 12. Gold-informed diagnostic references for the potential of structural information under a label-compatible edge composition, intervention-free test (group E11). All rows except N_0 use correct test labels and are not suitable for deployment.

Notation	Configuration	Macro-F1	vs. N_0	p_{adj}
N_0	matched baseline model	0.8179 ± 0.0124	—	—
A_{gold}	averaging on the sparse gold graph (oracle diagnostic)	0.8514 ± 0.0235	+0.0335	< 0.001
G_{gold}	controlled graph head on the sparse gold graph (oracle diagnostic)	0.8660 ± 0.0338	+0.0481	< 0.001
V_{gold}	voting by neighbors' gold labels (label-injection reference)	0.9669 ± 0.0033	+0.1490	< 0.001

The graph model on the gold graph exceeded N_0 by 0.0481; 95% confidence interval [0.0266; 0.0714], $p_{\text{adj}} < 0.001$. Averaging N_0 output distributions over the same gold-constructed graph (A_{gold}) also gave a gain of 0.0335, interval [0.0211; 0.0484], $p_{\text{adj}} < 0.001$. Thus, when gold types are used to determine links, the model is able to exploit the resulting label-compatible structure. The three rows serve different diagnostic purposes: gold-label voting (+0.1490, interval [0.1242; 0.1749]) directly injects neighbors' gold labels and is therefore a label-injection reference rather than a level attainable by a model using deployment-time inputs, whereas +0.0481 characterizes the ability of the trained head to exploit a fully label-compatible structure. Since gold labels already encode part of the answer in the graph, these results are not independent evidence of a signal obtainable from features available at deployment time. An advantage of the graph network over simple averaging was likewise not established: the difference of 0.0146 has interval [−0.0113; 0.0419], $p_{\text{adj}} = 0.276$.

The broader rule-based union contains 37,336 test edges with a label homogeneity of 0.3962; after filtering by gold labels, 14,794 edges remain with a homogeneity of 1.000, and the filtered graph exceeds the initial one by 0.0311 macro-F1 (interval [0.0203; 0.0429], $p_{\text{adj}} < 0.001$). For the untyped averaging operation used, this indicates the benefit of class-compatible neighbors, but does not establish that an edge connecting mentions of different types is inherently harmful: the context of a mention of a different type may be useful given a separate transformation for the corresponding link type. Moreover, filtering simultaneously changes the number of edges, the degrees, and the graph structure. The present experiment therefore does not separate errors of the edge-construction rules from the limitation of a model that combines different semantic links in a single channel. All main contrasts with confidence intervals are summarized in Figure 3.

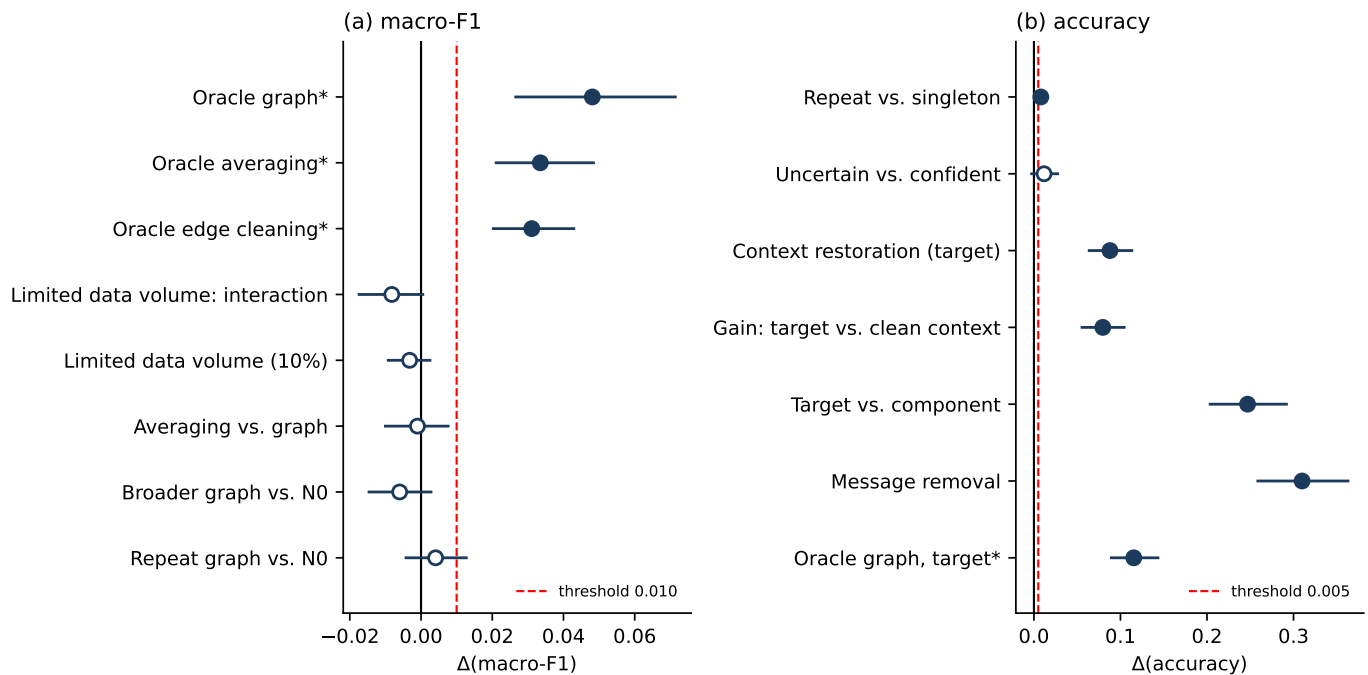


Figure 3. Main original second-stage contrasts with 95% confidence intervals: (a) macro-F1; (b) accuracy. A filled marker indicates the confidence interval excludes zero; an open marker indicates it includes zero; the dashed line marks the practical-significance threshold fixed before the second-stage analysis (0.010 for macro-F1, 0.005 for accuracy). Asterisks mark oracle diagnostics not suitable for deployment.

6.4. Controlled Context-Restoration Experiment

The controlled tasks illustrate the mechanism by which context transfer operates (Table 13, Figure 4). This experiment is central to testing RQ2.

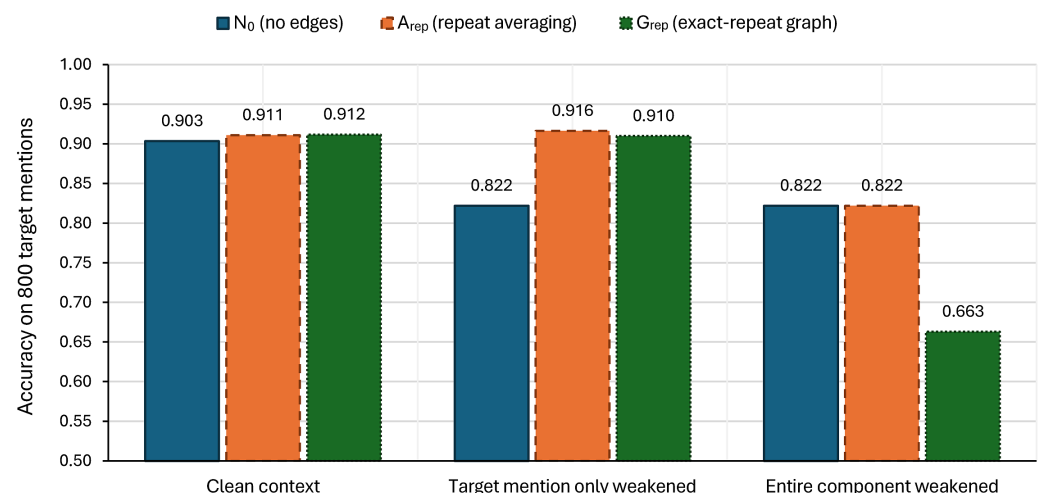


Figure 4. Accuracy on target mentions across three context regimes. In the whole-component regime, A_{rep} equals N_0 by construction because weakened neighbors are excluded from averaging.

When the external context is removed only for the target mention, the graph model, trained separately for the explicitly flagged weakening, increases accuracy by 0.0879 (95% CI [0.0639; 0.1131], $p_{adj} < 0.001$). The model approaches the clean-context performance level. When the same context is removed for the entire repeat component, accuracy drops to 0.6630, i.e., below the baseline model. The difference between these two contrasts (hereafter, the *interaction*) equals +0.2469, interval [0.2037; 0.2918], $p_{adj} < 0.001$. This

interaction indicates that, under this controlled protocol, the target mention's prediction depends on undamaged evidence in its neighbors: as soon as the neighbors also lose context, the gain does not merely vanish but turns into a loss. Removing all messages of an already-trained model reduces accuracy in the weakened-target regime by 0.3099 (interval $[0.2588; 0.3629]$, $p_{\text{adj}} < 0.001$), showing that the trained model depends on the availability of edges. The strictest regime, with the target mention replaced by the class-frequency distribution, shows the same pattern: with undamaged neighbors, the graph model achieves 0.9074 against 0.5300 for the baseline, and after weakening the entire component it drops to 0.0690. At the same time, simple averaging in the weakened-target regime reaches 0.9163, i.e., slightly exceeding the graph model's 0.9099. Recovery is therefore not a unique advantage of a complex neural architecture: the useful component in this artificial task is access to undamaged neighboring context.

Table 13. Accuracy on 800 target mentions (group E13). Values are the mean $\pm$ standard deviation across the 10 optimization seeds; each seed value is averaged over the 5 target-selection realizations. The clean-context row uses heads trained without weakening; the two weakened rows use heads trained under flagged target-only weakening.

Context Regime	N_0 (No Edges)	A_{rep} (Repeat Averaging)	G_{rep} (Exact-Repeat Graph)	$(G_{\text{rep}} - N_0)$
Clean context	0.9034 ± 0.0045	0.9109 ± 0.0077	0.9117 ± 0.0078	$+0.0083$
Target mention only weakened	0.8220 ± 0.0119	0.9163 ± 0.0031	0.9099 ± 0.0073	$+0.0879$
Entire component weakened	0.8220 ± 0.0119	$0.8220 \pm 0.0119^{\dagger}$	0.6630 ± 0.0286	-0.1590

[†] Equal to N_0 by construction: averaging excludes weakened neighbors, so no eligible neighbor remains (Section 5.3).

The matched indicator-control results are shown in Table 14. Holding $q_i = 0$ during training, validation, and testing reduces neither model to a test-time ablation: both heads are retrained under the same unflagged artificial weakening. The indicator-enabled and constant-indicator heads were trained in the same post-review run. This run used the same model and training settings, seeds, and target selections as the original E13 run, but a dedicated runner, a revised code base, a different operating system, and retrained local models. The clean-context rows of the two tables are not directly comparable: in Table 13, they come from heads trained without weakening, whereas in Table 14, they come from the weakening-trained heads evaluated on clean context. For the weakened rows, which share the training regime, the new run is close to the original (for example, a graph gain of 0.0913 versus 0.0879 under flagged target-only weakening). All indicator comparisons are made within the new run, and the original E13 values are reported unchanged.

Without an explicit degradation cue, G_{rep} exceeds the separately retrained N_0 by 0.0784 in the target-only condition (95% interval $[0.0544; 0.1021]$, $p_{\text{adj}} < 0.001$). The same-environment flagged replication gives a gain of 0.0913 (interval $[0.0676; 0.1157]$), close to the original E13 estimate of 0.0879. The indicator-dependence interaction, $(G_{\text{rep}} - N_0)_{q-\text{on}} - (G_{\text{rep}} - N_0)_{q-\text{off}}$, is 0.0129 with interval $[-0.0040; 0.0316]$ and $p_{\text{adj}} = 0.149$; dependence of the graph gain on the explicit cue is therefore not established.

When the entire repeat component is weakened without the cue, the graph contrast reverses to -0.1217 . The target-only minus whole-component interaction is $+0.2001$ (interval $[0.1638; 0.2385]$, $p_{\text{adj}} < 0.001$), preserving the diagnostic evidence that intact repeat context is the source of recovery. A_{rep} is a strong reference: it gains 0.0938 over N_0 in the unflagged target-only condition (interval $[0.0743; 0.1142]$), and it exceeds G_{rep} by 0.0154 (secondary reference contrast, interval $[0.0026; 0.0295]$). Because the target's repeat neighbors are intact in this condition, the mask-based neighbor exclusion described in Section 5.3 does not affect this comparison. Thus, the added control supports context transfer without q_i , not a

unique advantage of the learned graph head. The weakening remains artificial, and these results do not establish robustness to naturally occurring missing or corrupted context.

Table 14. Accuracy on the same 800 target mentions in the matched indicator control, with the degradation indicator enabled (q_i on) and held constant at zero ($q_i = 0$). Both regimes come from the same run, and all rows use heads trained under target-only weakening; the clean-context rows evaluate these heads with the target context intact. Values are the mean $\pm$ standard deviation across the 10 optimization seeds, each averaged over the 5 target-selection realizations (group E13-R). Confidence intervals and adjusted p -values for the primary contrasts are given in the text.

Context Regime	Indicator	N_0 (No Edges)	A_{rep} (Repeat Averaging)	G_{rep} (Exact-Repeat Graph)	$(G_{rep} - N_0)$
Clean context	q_i on	0.9105 ± 0.0052	0.9203 ± 0.0071	0.9283 ± 0.0049	$+0.0177$
	$q_i = 0$	0.9176 ± 0.0056	0.9259 ± 0.0056	0.9241 ± 0.0063	$+0.0066$
Target mention only weakened	q_i on	0.8206 ± 0.0146	0.9164 ± 0.0043	0.9119 ± 0.0083	$+0.0913$
	$q_i = 0$	0.8236 ± 0.0151	0.9174 ± 0.0053	0.9020 ± 0.0074	$+0.0784$
Entire component weakened	q_i on	0.8206 ± 0.0146	0.8206 ± 0.0146 [†]	0.6575 ± 0.0375	-0.1631
	$q_i = 0$	0.8236 ± 0.0151	0.8236 ± 0.0151 [†]	0.7019 ± 0.0236	-0.1217

[†] Equal to N_0 by construction: averaging excludes weakened neighbors, so no eligible neighbor remains (Section 5.3).

6.5. Effect of the Specific Topology

The importance of the specific link structure was tested by shuffling edges while preserving node degrees. This intervention changes, on average, 56.31% of the edges of the repeat graph. On the intervention-free test, the original graph exceeds the shuffled one by only 0.0011 ± 0.0008 macro-F1 and by 0.0004 ± 0.0006 accuracy, whereas in the weakened-target-mention regime, the difference grows to 0.1655 ± 0.0200 accuracy (mean and standard deviation across the 10 optimization seeds, each averaged over the 5 topology realizations). Thus, the effect of the specific links is negligible in the aggregate but becomes substantially larger once the local evidence of the target mention has been weakened. Unlike complete message removal, which belongs to family E13, the shuffling contrasts were not included in the predefined hypothesis families and are therefore reported descriptively, without bootstrap intervals or corrected p -values.

This control is conservative: label homogeneity after shuffling decreases only from 0.9970 to 0.8510. For the graph built from gold labels, the intervention is stronger—90.81% of edges change, homogeneity drops from 1.000 to 0.4361, and macro-F1 decreases by 0.0741 ± 0.0429 ; complete removal of these edges reduces it by 0.1269 ± 0.0404 . This result indicates the model's sensitivity to the gold-defined structure, but, because correct types are used in the graph construction itself, does not establish the availability of such a signal at deployment time.

6.6. Effect of Training-Data Volume

The dependence of the effect on training-data volume is shown in Table 15 and Figure 5.

Table 15. Effect of training-data volume on macro-F1.

Fraction of Training Documents	Documents	Mentions	$\Delta (G_{rep} - N_0)$, Macro-F1
10%	34	1406	-0.0032
25%	84	3298	-0.0051
50%	167	6406	$+0.0137$
100%	333	12,799	$+0.0041$

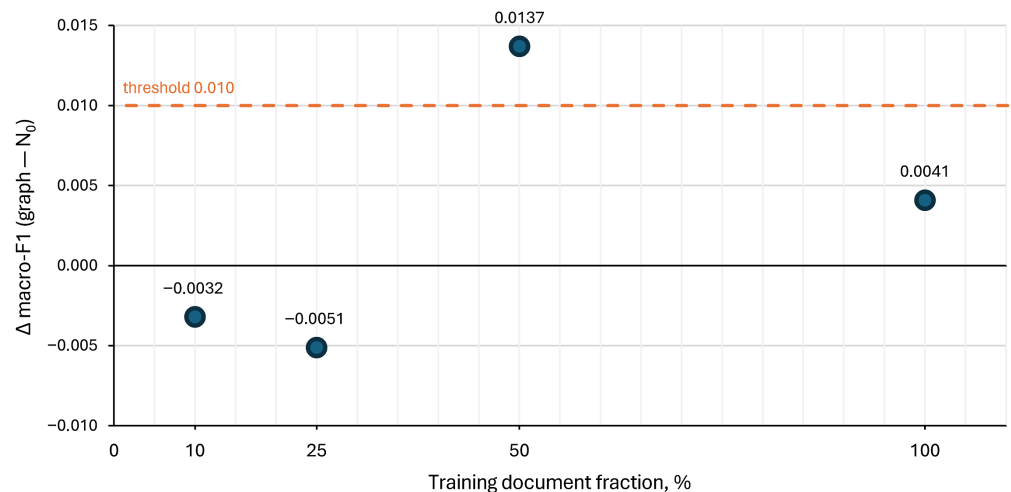


Figure 5. Effect of graph-based refinement on macro-F1 across training-data fractions.

Reducing the training set alone does not strengthen the benefit of the graph. For 10% of the data, the contrast equals -0.0032 , interval $[-0.0092; 0.0025]$, $p_{\text{adj}} = 0.281$. The interaction of the average effect for 10% and 25% versus the full set equals -0.0082 , interval $[-0.0174; 0.0005]$, $p_{\text{adj}} = 0.140$. The observed dependence is non-monotonic, so it provides no basis for claiming either a systematic increase in the effect as the training set shrinks or an optimum at 50% of the training data. Weak evidence for a specific mention and an overall scarcity of training labels are different regimes: the graph network itself requires sufficient data for training.

6.7. Distribution of the Effect Across Mention Groups

The distribution of the effect across mention groups is shown in Table 16.

Table 16. Descriptive analysis of mention groups: accuracy contrasts, mean $\pm$ standard deviation over 10 training runs (group E15).

Mention Group	Support	$\Delta (G_{\text{rep}} - N_0)$, Accuracy
Repeated	2385	$+0.0083 \pm 0.0056$
Singleton	4546	$+0.0003 \pm 0.0063$
Uncertain	1771	$+0.0117 \pm 0.0226$
Confident	5160	$+0.0000 \pm 0.0001$
Uncertain and repeated simultaneously	491	$+0.0399 \pm 0.0270$
Training-ambiguous	82	$+0.0037 \pm 0.0215$

The values in Table 16 are reported primarily for descriptive purposes; controlled second-stage inference is limited to the two comparisons fixed for family E15. The interaction of repeated versus singleton mentions equals $+0.0080$, interval $[0.0029; 0.0129]$, $p_{\text{adj}} = 0.004$. This is consistent with the graph construction: the exact-repeat graph can help only when a corresponding edge exists. For uncertain mentions, the sign is likewise positive, but the interaction with confident mentions is not statistically established: $+0.0117$, interval $[-0.0026; 0.0274]$, $p_{\text{adj}} = 0.122$. The intersection of uncertain and repeated mentions has the largest point estimate ($+0.0399$), but this estimate is descriptive, and the group of training-ambiguous forms (82 mentions) is too small to support a reliable conclusion.

6.8. Summary of Research-Question Verification

The correspondence between the research questions, the experiments, the observed effects, and the conclusions is summarized in Table 17.

Table 17. Summary of findings for each research question.

Question	What Was Tested	Experiment	Effect	Conclusion
RQ1	whether the graph improves overall quality	intervention-free test, E12	+0.0041 macro-F1, [−0.0042;0.0127], $p_{\text{adj}} = 0.651$	benefit not established; meaningful benefit not excluded
RQ2	whether the graph can recover lost context	flagged target-only weakening, E13	+0.0879 accuracy, [0.0639;0.1131], $p_{\text{adj}} < 0.001$	supported in the controlled regime
RQ2	whether recovery persists without q_i	constant-indicator target weakening, E13-R	+0.0784, [0.0544;0.1021], $p_{\text{adj}} < 0.001$	yes, under unflagged artificial weakening
RQ2	whether recovery depends on an undamaged neighbor	“target vs. component” interaction, E13	+0.2469, [0.2037;0.2918], $p_{\text{adj}} < 0.001$	diagnostic dependence under the tested weakening; not a general necessity
RQ3	whether the specific link structure matters	topology shuffling, E13	+0.1655 accuracy *	yes, under weakened context
RQ3	whether edge composition is a limitation	oracle cleaning, E11–E12	+0.0311 macro-F1, [0.0203;0.0429], $p_{\text{adj}} < 0.001$	yes, under oracle cleaning
RQ3	whether lemma edges improve over exact repeats	morphology control, E12-M	−0.0033, [−0.0135;0.0055], $p_{\text{adj}} = 0.954$	no advantage established
RQ3	whether a complex graph head is needed	A_{rep} vs. G_{rep} , E12	−0.0010, [−0.0100;0.0076], $p_{\text{adj}} = 0.825$	no advantage established
RQ4	whether the effect is stronger for repeated mentions	repeated vs. singleton, E15	+0.0080, [0.0029;0.0129], $p_{\text{adj}} = 0.004$	yes
RQ4	whether the effect is stronger for uncertain mentions	uncertain vs. confident, E15	+0.0117, [−0.0026;0.0274], $p_{\text{adj}} = 0.122$	not established
RQ4	whether the “uncertain and repeated” group is promising	subgroup analysis, E15	+0.0399 (491 mentions)	descriptively promising
—	whether a smaller training set strengthens the graph’s benefit	10–100% training data, E14	non-monotonic; interaction −0.0082, [−0.0174;0.0005], $p_{\text{adj}} = 0.140$	not established

* Descriptive counterfactual contrast outside the predefined hypothesis families; a corrected p -value was not computed.

7. Discussion

7.1. Overall Effect, Context Recovery, and Edge Rules

On the intervention-free test, an overall advantage of the exact-repeat graph over the matched no-message head was not established: the point gain was 0.0041 macro-F1 and its interval crossed zero while extending above the 0.010 practical threshold. The morphology control suggests that incomplete exact coverage is not the only constraint. Lemmatization raised coverage from 34.41% to 42.07% and retained 0.9959 label homogeneity, yet G_{lem} did not establish an improvement over G_{rep} (−0.0033, 95% CI [−0.0135;0.0055]) or a benefit over N_0 (+0.0030, [−0.0053;0.0117]; the upper bound still admits an effect above 0.010). The experiment does not identify why; plausible explanations are that local evidence is already sufficient for many newly linked mentions, that analyzer errors introduce false links, or that the aggregation rule cannot exploit the added relations.

The controlled interventions isolate a narrower positive mechanism. Under target-only weakening, $G_{\text{rep}} - N_0$ is +0.0879 accuracy when q_i is available and +0.0784 (95% CI

[0.0544;0.1021]) when q_i is fixed to zero throughout training and evaluation. When the whole repeat component is weakened, the corresponding graph contrasts become negative, producing target-versus-component interactions of +0.2469 and +0.2001, respectively. Detectable dependence on the explicit indicator was not established. These results support transfer from an intact neighbor under synthetic corruption, but holding q_i constant does not transform that intervention into naturally occurring noise, and component weakening remains a diagnostic counterfactual. All context-recovery evidence (E13 and E13-R) was obtained with exact normalized repeats; whether comparable transfer occurs between morphologically varying mentions was not tested, because the lemma rule was evaluated only on intervention-free data.

Topology and aggregation controls point in the same direction. Degree-preserving shuffling changes little on the clean test (+0.0011 macro-F1 for the original over shuffled topology) but has a much larger effect when target context is weakened (+0.1655 accuracy). Simple averaging is also competitive: $A_{rep} - G_{rep} = -0.0010$ on the original intervention-free comparison, and in the no-indicator target control A_{rep} reaches 0.9174 accuracy versus 0.9020 for G_{rep} . Thus, the study does not establish an advantage of the learned graph head over a simple propagation baseline. Equal nominal parameter counts likewise do not imply equal effective capacity, because the neighbor projection and gate are inactive in N_0 ; the topology controls are needed to complement that comparison.

The precision–coverage trade-off remains central for Ukrainian. The exact rule misses the inflectional pair “Kateryny Kobchenko” and “Kateryna Kobchenko” in Figure 2; the manual audit found inflectional or shortened variants in 24 of 29 clear missed-counterpart cases. The lemma rule recovers many such forms, but its context-free first-ranked analysis can also conflate ambiguous names or common nouns. Conversely, exact strings such as “Kyiv” and “Ukraine” can denote different types, which motivates document-local rather than corpus-wide matching. Neither rule performs coreference resolution, and the corpus does not contain entity identifiers with which to measure true link recall.

7.2. Which Mentions Benefit Most

The effect is concentrated among mentions with a corresponding edge: the interaction of repeated versus singleton mentions equals +0.0080 ($p_{adj} = 0.004$), whereas for singleton mentions, the gain is practically absent (+0.0003). The intersection of repetition and high local uncertainty has the largest descriptive estimate (+0.0399 on 491 mentions), but the interaction of uncertain versus confident mentions was not statistically established ($p_{adj} = 0.122$). This result should be regarded as hypothesis-generating rather than as grounds for a routing rule; any such strategy should be evaluated prospectively in a separate experiment with a predefined threshold, reporting coverage, accuracy on the routed subset, calibration, and global macro-F1.

7.3. Comparison with Prior Research

A comparison with prior approaches is given in Table 18. Because the task formulations differ in language and corpus, and because the present study assumes known mention boundaries, absolute F1-values are not directly comparable. Gains over each work’s own baseline model are therefore compared; such a comparison should be regarded as contextual rather than as a direct system-level comparison.

In terms of point-estimate magnitude, the gain in the present study (+0.41 points of macro-F1) is of the same order as the gains reported in prior studies (from +0.32 to +1.00 points). However, direct comparison of gain magnitudes is limited by differences in corpora, languages, baseline models, and experimental settings. The contribution here is not a new matching rule, graph layer, oracle diagnostic, or topology perturbation; these are

mature tools. It is the controlled combination of those tools to expose when the apparent link effect survives matched retraining, altered topology, simpler propagation, morphology-aware coverage, and context interventions in Ukrainian. In the work of P. Chen and co-authors [21], the baseline model has neither a graph-neural-network layer nor a fusion layer, so the measured difference also includes the contribution of additional parameters. The matched-parameterization comparison in the present study controls nominal head size and non-structural inputs, but not effective prediction capacity, because the neighbor branch is inactive in N_0 . It shows that the observed point estimate remains statistically uncertain: the 95% confidence interval $[-0.42; +1.27]$ points cross zero while still including gains above the practical threshold. This does not refute the results obtained by A. Akbik and co-authors [17], S. Schweter and A. Akbik [18], T. Gui and co-authors [19], A. Hu and co-authors [20], and P. Chen and co-authors [21], but indicates that isolating the contribution of links requires complementary controls.

Table 18. Comparison with prior work using document-level links for NER.

Work	Corpus	Gain, Pts.	Matched Baseline	Random-Topology Check
Akbik et al. [17]	CoNLL-03 EN	+0.32	no	no
Schweter, Akbik [18]	CoNLL-03 EN	+0.85	yes	no
Gui et al. [19]	CoNLL-2003	+0.92	no	no
Hu et al. [20]	CoNLL-2003	+1.00	yes	no
Chen et al. [21]	AnatEM	+0.88	no	no
Chen et al. [21]	MARS	+0.62	no	no
This work	NER-UK 2.0	+0.41	yes	yes

In addition, oracle filtering and shuffling indicate sensitivity of the result to the composition and placement of links, but the present single-channel experiment does not separate the accuracy of edge estimation from the model's ability to distinguish their semantics. Further verification should therefore jointly evaluate improved link rules and typed aggregation mechanisms.

7.4. Practical Implications

From a practical standpoint, the results do not support the unconditional use of graph-based refinement for all mentions. A selective strategy that uses graph information primarily for repeated mentions with insufficient local context is a plausible candidate but was not evaluated here. In an end-to-end system, the evaluated module could only operate as a second-stage component after mention detection; upstream missed, spurious, or incorrectly bounded spans would propagate into both node representations and graph construction. At the same time, a routing rule based on uncertainty requires a separate prospective check, since in the present study, an advantage of uncertain over confident mentions was not statistically established. Because simple averaging was competitive, it should serve as the baseline implementation of context transfer in a practical system, while a trained graph head should be treated as a variant requiring separate justification.

7.5. Limitations of the Study

Limitations of the problem statement. The formulation assumes known mention boundaries, so the reported metrics do not represent end-to-end NER performance: in a real system, an upstream boundary-detection module may introduce errors. A frozen XLM-RoBERTa-base deliberately gives all compared heads identical local representations and avoids confounding link effects with encoder adaptation, but it does not establish how the graph would behave if the encoder were fine-tuned jointly with it. The context-

weakening regimes are artificial interventions. The original E13 models explicitly receive the indicator q_i ; the post-review constant-indicator control shows that the target-restoration effect persists without that cue, but still does not evaluate robustness to naturally occurring noise. The current evidence therefore supports context transfer under synthetic corruption only; it does not establish robustness to OCR errors, automatic speech recognition (ASR) errors, headlines, or abbreviated texts, which are named as future work in Section 8 but were not tested here. Component weakening also remains a diagnostic counterfactual rather than a deployment condition.

Limitations of the graph. Exact surface-form repetition is only an approximation of entity identity: identical spelling sometimes denotes different objects, and inflected, abbreviated, or otherwise variant forms can denote the same entity without matching spelling. Exact-repeat participation is 34.41% overall and varies markedly by class. The lemma rule raises coverage to 42.07%, but a context-free first-ranked morphological analysis can mislemmatize proper names or merge ambiguous forms; its trained graph did not establish an improvement over exact repetition. The morphology control and the main experiments also differ in scope: E12-M compares exact and lemma edges only on intervention-free data, with heads retrained in a separate run, whereas the context-recovery experiments (E13 and E13-R) use exact repeats only. Its results therefore concern coverage and intervention-free performance, not context recovery between morphological variants. Moreover, the precision–coverage trade-off of the lemma rule remains only partially quantified: the added coverage is measured, but the precision of the 2767 added links was not audited, and post-construction label homogeneity is not a measure of same-entity precision. This limits the interpretation of the negative lemma result, which cannot be attributed specifically to redundant local evidence, analyzer errors, or the aggregation rule. Shared-sentence and proximity edges likewise express co-occurrence rather than a semantic relation and can join unrelated or differently typed mentions; they must not be interpreted as coreference or general semantic-relation edges. Because the corpus lacks entity identifiers or coreference chains, neither coverage value can be converted into true link recall; moreover, the 100-mention, single-auditor manual sample is descriptive rather than a gold-standard evaluation. Oracle graphs use correct test types to determine edges or messages, encode part of the answer, and are not independent evidence of an inter-mention signal available at deployment time. Label homogeneity is an indicator of class compatibility for the untyped aggregation used, not a universal definition of edge quality; because of the single adjacency channel, the experiment does not separate errors of the link rules from limitations of the aggregation method. Edge shuffling preserves the document and the degree of every node, yet afterward, the homogeneity of the repeat graph remains high (0.8510), so it is a conservative control rather than a fully label-neutral random topology.

Limitations of generalizability. The low-data analysis relies on a single fixed sequence of subsamples and does not capture variability across alternative small training samples. The exploratory and fixed second-stage analyses used the same official test set, and both post-review controls used that split again; therefore, the reported statistical analysis is controlled evidence within a previously inspected dataset and cannot substitute for independent external replication. Classes DOC (40 mentions) and TIME (10 mentions) have too few test mentions, so conclusions about them require separate per-class confidence intervals and confusion matrices. The nominal parameter counts of N_0 and the graph heads are equal, but their effective prediction capacities are not; the comparison must be interpreted jointly with the topology controls. A formal equivalence test between simple averaging and the graph head was not performed.

8. Conclusions

This study presented a controlled empirical framework for evaluating the contribution of document-level links to named-entity type classification in Ukrainian texts. Its contribution is the integrated evaluation design—a matched-parameterization baseline, controlled topologies, oracle diagnostics, simple propagation references, morphology-aware edges, and context interventions—rather than a fundamentally new graph architecture. The baseline matches the graph head’s parameterization and nominal parameter count, but not its effective prediction capacity; topology controls therefore provide a complementary test of link placement. The framework was evaluated on NER-UK 2.0 using frozen XLM-RoBERTa-base representations and known mention boundaries.

The intervention-free evaluation did not establish an overall advantage of graph-based refinement over the matched-parameterization baseline. The point estimate for the exact-repeat graph was $+0.0041$ macro-F1 ($p_{\text{adj}} = 0.651$), below the predefined practical-significance threshold, but the confidence interval still included effects above that threshold. In contrast, under explicitly flagged weakening of the local context of the target mention, graph-based refinement increased target accuracy by 0.0879 . The post-review control retained a gain of 0.0784 (95% CI $[0.0544; 0.1021]$) when the degradation indicator was held constant for both heads, and detectable dependence on the indicator was not established. Under whole-component weakening without the indicator, the advantage reversed to -0.1217 , producing an interaction of $+0.2001$ between target-only and component weakening (95% CI $[0.1638; 0.2385]$). These results support document-level context transfer when a target mention lacks sufficient local evidence while a linked repeated mention retains informative context; they do not establish robustness to naturally occurring noise.

The morphology control raised graph coverage from 34.41% to 42.07% , but did not establish an improvement of G_{lem} over G_{rep} (-0.0033 , 95% CI $[-0.0135; 0.0055]$) or over N_0 ($+0.0030$, $[-0.0053; 0.0117]$). The experiments further showed that the exact-repeat effect is significantly stronger for repeated than for singleton mentions ($p_{\text{adj}} = 0.004$), whereas a stronger effect for uncertain than for confident mentions was not statistically established ($p_{\text{adj}} = 0.122$). Oracle label-compatible graphs demonstrated additional potential for improvement, but this result is a gold-informed diagnostic reference, not a deployment result, because such links depend on gold test labels. Moreover, simple averaging remained competitive and outperformed the trained graph head in the no-indicator target control, so the experiments do not establish an advantage of a more complex graph architecture over simpler propagation.

The principal conclusion is therefore that graph-based refinement should not be regarded as a universal replacement for local named-entity classification. Rather, it is a conditional mechanism for recovering missing contextual information whose effectiveness depends on the availability of a relevant linked mention and on the reliability of the link connecting it to the target. This finding shifts the main design question from whether a graph should be added to an NER model to when document-level information should be propagated and which links are sufficiently reliable to support such propagation.

The conclusions are subject to several limitations. Mention boundaries were assumed to be known, the encoder was frozen, and the context-weakening scenarios were artificial even in the constant-indicator control. Exact and lemma edge coverage are not link-recall estimates because the corpus has no entity identifiers or coreference chains; the context-free morphological analyzer can also introduce false matches. The fixed second-stage evaluation and both post-review controls reused a previously inspected test split. Consequently, the results are controlled but not independently confirmatory, and validation on an untouched corpus is required before the observed effects can be considered generalizable.

Future research should therefore focus on context-sensitive morphology, entity linking, or learned inter-mention links; explicitly evaluate their precision–coverage trade-off; and develop selective routing mechanisms that activate document-level context only when local evidence is insufficient. Further experiments should compare learned refinement with simpler propagation, distinguish relation types through separate transformations or weights, and evaluate context recovery under naturally degraded conditions such as OCR errors, automatic speech transcription, headlines, and abbreviated texts. Joint encoder–graph fine-tuning is another relevant extension, but it requires controls that separate link effects from encoder adaptation. Corpora containing entity identifiers or coreference annotation would enable direct evaluation beyond graph participation and label homogeneity.

Author Contributions: Conceptualization, O.H.; methodology, O.H. and V.S.; software, O.H.; validation, O.H., V.S. and D.G.; formal analysis, O.H.; investigation, O.H.; resources, O.H.; data curation, O.H.; writing—original draft preparation, O.H. and V.S.; writing—review and editing, O.H., V.S. and D.G.; visualization, O.H. and D.G.; supervision, V.S.; project administration, V.S.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: The APC of the journal was supported by the Faculty of Electrical and Computer Engineering, Cracow University of Technology and the Ministry of Science and Higher Education, Republic of Poland (grant no. E-1/2026).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code, experimental configurations, and tests required to reproduce the reported experiments, including both post-review controls, are publicly available at <https://github.com/KepAlex503/ner-gnn-fusion-tests> (accessed on 1 Month 2026) (tag revision-1). Trained model checkpoints, predictions, metric files, and bootstrap outputs are available from the corresponding author upon reasonable request. The NER-UK 2.0 corpus used for training and evaluation is available under the CC BY-NC-SA 4.0 license from its original source [11].

Acknowledgments: During the preparation of this manuscript, the authors used generative AI tools (Claude Opus 5, Claude Opus 5.5, Claude Sonnet 5; Anthropic, San Francisco, CA, USA) for English proofreading and grammatical editing, for LaTeX formatting, and as an aid in the manual review of the 100 audited mentions described in Section 5; all audit decisions were verified by the authors. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NER	Named-Entity Recognition
GNN	Graph Neural Network
CI	Confidence Interval
MLP	Multilayer Perceptron
RQ	Research Question

References

1. Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. *Linguisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
2. Goyal, A.; Gupta, V.; Kumar, M. Recent Named Entity Recognition and Classification Techniques: A Systematic Review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]

3. Jehangir, B.; Radhakrishnan, S.; Agarwal, R. A Survey on Named Entity Recognition—Datasets, Tools, and Methodologies. *Nat. Lang. Process. J.* **2023**, *3*, 100017. [\[CrossRef\]](#)
4. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 260–270. [\[CrossRef\]](#)
5. Akbik, A.; Blythe, D.; Vollgraf, R. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1638–1649.
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [\[CrossRef\]](#)
7. Seow, W.L.; Chaturvedi, I.; Hogarth, A.; Mao, R.; Cambria, E. A Review of Named Entity Recognition: From Learning Methods to Modelling Paradigms and Tasks. *Artif. Intell. Rev.* **2025**, *58*, 315. [\[CrossRef\]](#)
8. Hu, W.; He, L.; Ma, H.; Wang, K.; Xiao, J. KGNER: Improving Chinese Named Entity Recognition by BERT Infused with the Knowledge Graph. *Appl. Sci.* **2022**, *12*, 7702. [\[CrossRef\]](#)
9. Sun, W.; Liu, S.; Liu, Y.; Kong, L.; Jian, Z. Named Entity Recognition Networks Based on Syntactically Constrained Attention. *Appl. Sci.* **2023**, *13*, 3993. [\[CrossRef\]](#)
10. Katrenko, S.; Adriaans, P. Named Entity Recognition for Ukrainian: A Resource-Light Approach. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, Prague, Czech Republic, 29 June 2007; pp. 88–93.
11. Chaplynskyi, D.; Romanyshyn, M. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian. In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, Turin, Italy, 25 May 2024; pp. 23–29. [\[CrossRef\]](#)
12. Radchenko, V.; Drushchak, N. Improving Named Entity Recognition for Low-Resource Languages Using Large Language Models: A Ukrainian Case Study. In Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), Vienna, Austria, 31 July–1 August 2025; pp. 27–35. [\[CrossRef\]](#)
13. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451. [\[CrossRef\]](#)
14. Wang, H.; Zhou, L.; Duan, J.; He, L. Cross-Lingual Named Entity Recognition Based on Attention and Adversarial Training. *Appl. Sci.* **2023**, *13*, 2548. [\[CrossRef\]](#)
15. Li, Q.; Li, F.; Li, S.; Li, X.; Liu, K.; Liu, Q.; Dong, P. Improving Entity Linking by Introducing Knowledge Graph Structure Information. *Appl. Sci.* **2022**, *12*, 2702. [\[CrossRef\]](#)
16. Li, H.; Li, C.; Sun, Z.; Zhu, H. Entity Linking Model Based on Cascading Attention and Dynamic Graph. *Electronics* **2024**, *13*, 3845. [\[CrossRef\]](#)
17. Akbik, A.; Bergmann, T.; Vollgraf, R. Pooled Contextualized Embeddings for Named Entity Recognition. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 724–728. [\[CrossRef\]](#)
18. Schweter, S.; Akbik, A. FLERT: Document-Level Features for Named Entity Recognition. *arXiv* **2020**, arXiv:2011.06993. [\[CrossRef\]](#)
19. Gui, T.; Ye, J.; Zhang, Q.; Zhou, Y.; Gong, Y.; Huang, X. Leveraging Document-Level Label Consistency for Named Entity Recognition. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization: Darmstadt, Germany, 2020; pp. 3976–3982. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Hu, A.; Dou, Z.; Nie, J.Y.; Wen, J.R. Leveraging Multi-Token Entities in Document-Level Named Entity Recognition. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 7961–7968. [\[CrossRef\]](#)
21. Chen, P.; Ding, H.; Araki, J.; Huang, R. Explicitly Capturing Relations between Entity Mentions via Graph Neural Networks for Domain-specific Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 1–6 August 2021; pp. 735–742. [\[CrossRef\]](#)
22. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1064–1074. [\[CrossRef\]](#)
23. Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2145–2158.
24. Dai, Z.; Fei, H.; Li, P. Coreference Aware Representation Learning for Neural Named Entity Recognition. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization: Darmstadt, Germany, 2019; pp. 4946–4953. [\[CrossRef\]](#) [\[PubMed\]](#)

25. Luo, Y.; Xiao, F.; Zhao, H. Hierarchical Contextualized Representation for Named Entity Recognition. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8441–8448. [[CrossRef](#)]
26. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
27. Hamilton, W.L.; Ying, Z.; Leskovec, J. Inductive Representation Learning on Large Graphs. In *Proceedings of the Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 1024–1034.
28. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
29. Qian, Y.; Santus, E.; Jin, Z.; Guo, J.; Barzilay, R. GraphIE: A Graph-Based Framework for Information Extraction. 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), Minneapolis, MN, USA, 2–7 June 2019; pp. 751–761. [[CrossRef](#)]
30. Chen, Y.; Shi, B. Enhanced Heterogeneous Graph Attention Network with a Novel Multilabel Focal Loss for Document-Level Relation Extraction. *Entropy* **2024**, *26*, 210. [[CrossRef](#)] [[PubMed](#)]
31. pymorphy3 Contributors. pymorphy3: Morphological Analyzer for Python. Version 2.0.6. 2025. Available online: <https://pypi.org/project/pymorphy3/> (accessed on 1 October 2026).
32. pymorphy3 Contributors. Ukrainian Morphological Dictionaries for pymorphy3. Version 2.4.1.1.1663094765. 2022. Available online: <https://pypi.org/project/pymorphy3-dicts-uk/> (accessed on 1 October 2026).
33. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
34. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
35. Reimers, N.; Gurevych, I. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 338–348. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.