

Who Is the Story About?

Protagonist Entity Recognition in News

Jorge Gabín^{1,2*}, M. Eduardo Ares¹, Javier Parapar²

¹Linknovate Science, A Coruña, Spain

²IRLab, CITIC, Computer Science Department, University of A Coruña, Spain
{jorge.gabin,javier.parapar}@udc.es, eduardo@linknovate.com

Abstract

News articles often reference numerous organizations, but traditional Named Entity Recognition (NER) treats all mentions equally, obscuring which entities genuinely drive the narrative. This limits downstream tasks that rely on understanding event salience, influence, or narrative focus. We introduce *Protagonist Entity Recognition* (PER), a task that identifies the organizations that anchor a news story and shape its main developments. To validate PER, we compare the predictions of Large Language Models (LLMs) against annotations from four expert annotators over a gold corpus, establishing both inter-annotator consistency and human-LLM agreement. Leveraging these findings, we use state-of-the-art LLMs to automatically label large-scale news collections through NER-guided prompting, generating scalable, high-quality supervision. We then evaluate whether other LLMs, given reduced context and without explicit candidate guidance, can still infer the correct protagonists. Our results demonstrate that PER is a feasible and meaningful extension to narrative-centered information extraction, and that guided LLMs can approximate human judgments of narrative importance at scale.

1 Introduction

News articles are fundamentally narratives about actors and actions: they describe who did what, why it matters, and what consequences follow. While modern Information Extraction (IE) systems have made substantial progress in identifying entities and linking them to canonical knowledge bases (Grishman, 1997), they generally treat all entity

mentions as equally important. In practice, however, some entities play a central narrative role, while others are merely contextual. For example, a financial report on a company’s quarterly performance may reference competitors, regulators, and partners, yet only one or two organizations truly anchor the story. Traditional Named Entity Recognition (NER) captures all such mentions without distinguishing their narrative roles (Tjong Kim Sang and De Meulder, 2003), leaving downstream applications such as media monitoring, event analysis, or knowledge-graph construction to filter large amounts of irrelevant information. What is often needed instead is an answer to a more interpretive question: *who is the protagonist of this story?*

We refer to this challenge as *Protagonist Entity Recognition (PER)* (Table 1). Unlike NER, which is primarily lexical and localized, PER is discourse-aware and interpretive. Determining a protagonist requires understanding how narrative emphasis is distributed across the document, integrating signals from headlines, leads, quotes, and event descriptions. Identifying protagonists therefore bridges the gap between surface-level mention detection and narrative-centered comprehension. While related work on entity and event salience (Dunietz and Gillick, 2014; Bhowmik et al., 2024) aims to estimate importance, no existing benchmark directly evaluates which organizations are central to a news article’s narrative arc.

A major obstacle to progress is the absence of a task formalization and accompanying data. It is not clear, for instance, whether humans consistently agree on which entities in a news story are protagonists, nor how well current Large Language Models (LLMs)

*Corresponding author: jorge.gabin@udc.es

Example Article Excerpt:

“TechCorp announced record quarterly earnings on Tuesday, citing strong demand in Asian markets. The report also mentioned increased regulatory scrutiny from the European Commission and rising competition from GlobalSoft and InfoDynamics. Analysts expect TechCorp to expand further in the coming year.”

Entities Recognized by NER:

TechCorp, European Commission, GlobalSoft, InfoDynamics

Protagonist Entities:

TechCorp

Table 1: Illustrative example showing that NER captures multiple entity mentions, while PER identifies which organization anchors the narrative.

can approximate such judgments. To address this, we first conducted a controlled annotation study in which four expert annotators independently labeled the protagonist organizations in a set of 50 news articles. This gold-standard corpus enabled us to measure inter-annotator agreement and assess the reliability of the task itself. We then compared human annotations to LLM predictions, evaluating whether models can match the nuance of human narrative interpretation rather than simply identifying frequent or headline-mentioned entities.

Our findings indicate that, when appropriately guided (Ouyang et al., 2022), LLMs achieve levels of agreement with human annotators comparable to the agreement levels among the annotators themselves. This result suggests that LLMs can serve as effective and reliable annotators for scaling the task beyond manually curated datasets. Building on this insight, we constructed a large-scale PER corpus by prompting state-of-the-art LLMs with NER-derived candidate entities, generating supervision signals that reflect narrative centrality rather than mere mention frequency. We further evaluated whether other LLMs, provided with substantially less context and without explicit candidate lists, can still identify protagonists purely from the document narrative.

This two-stage pipeline—manual validation followed by automatic expansion—contributes both a formal task definition and practical resources for studying narrative centrality in news. By demonstrating that PER is both

learnable and scalable, we highlight its potential to improve a range of downstream applications that depend on understanding which actors truly matter in a story. More broadly, our work argues that information extraction should evolve from recognizing *who is mentioned* to understanding *who the story is about*, moving automated text analysis closer to genuine narrative comprehension.

Our Contribution. We build on these lines of research by (1) formalizing protagonist identification as a document-level discourse interpretation task, (2) establishing a human-annotated benchmark to assess reliability, and (3) evaluating whether contemporary LLMs can both replicate human judgments and scale labeling to larger corpora. In doing so, we connect human-centered annotation methodology with LLM-based corpus construction, and highlight when and how LLMs can approximate narrative understanding.

2 Related Work

Named Entity Recognition and Entity Salience. Traditional NER focuses on identifying and categorizing entity mentions (Tjong Kim Sang and De Meulder, 2003), but it makes no inference about the narrative function of those entities. Entity salience modeling attempts to determine which entities are most central to a document’s meaning (Gamon et al., 2013; Liao and Veeramachaneni, 2009; Gillenwater et al., 2012). Salience is typically estimated through features such as mention frequency, syntactic prominence, coreference chains, or discourse structure (Clark and Manning, 2015). However, salience does not necessarily imply protagonism: an entity may be frequently mentioned yet play a peripheral or reactive role. In contrast, *Protagonist Entity Recognition* seeks to identify entities that drive the narrative, shape event progression, and define the story’s thematic focus.

Event Extraction and Semantic Role Labeling. Event extraction and semantic role labeling aim to determine *who did what to whom* in specific event frames (Xiang and Wang, 2019; Marcus et al., 1993). While these approaches support structured repre-

sentations of actions, they are event-centric rather than narrative-centric. A protagonist may appear across multiple events, episodes, or discourse segments, contributing to narrative continuity beyond any single predicate or event frame. This distinction aligns PER more closely with discourse-level analysis than with isolated event role identification.

Narrative Understanding and Summarization. Work on narrative structure, character modeling, and story summarization has highlighted the importance of identifying central actors in narrative texts (Elsner, 2012; Guan et al., 2020). However, most such work focuses on literary narratives or long-form stories, where protagonist roles are comparatively explicit. News narratives are compact, heterogeneous, and fact-driven, making protagonism less stylistically marked and more dependent on pragmatic, contextual, and discourse cues. PER thus extends narrative reasoning into the domain of real-world informational text.

LLMs for Annotation and Weak Supervision. Recent work has shown that LLMs can act as effective zero-shot or few-shot annotators for a wide range of linguistic tasks (Gilardi et al., 2023; Zhao et al., 2023). LLMs have also been used to generate weak supervision (Meng et al., 2022), enabling corpus expansion without manual labeling cost. However, only limited research has examined whether LLMs can reliably approximate *interpretive* judgments—such as narrative centrality—that require global document reasoning rather than local classification.

3 Annotating Protagonists

Establishing a reliable annotation framework is essential for studying PER. Since no prior dataset captures this notion, we first defined an annotation protocol and conducted a controlled human study to evaluate whether the concept of a *protagonist organisation*—the entity that anchors the main narrative of a news article—can be consistently identified. Once the task proved interpretable and sufficiently reliable, we extended it to LLMs to examine whether they could emulate human judgement and scale corpus construction.

3.1 Manual Annotation

We sampled fifty articles from the Finer-ORD corpus (Shah et al., 2024), which contains financial and organisational news. An off-the-shelf NER system (Tjong Kim Sang and De Meulder, 2003) was used to extract candidate organisational mentions; annotators could add missing entities when warranted. Four professional annotators, anonymised as A1–A4, were provided with concise guidelines framing protagonists as organisations whose actions, decisions, or state form the central focus of the article. Annotators were encouraged to treat the headline and lead paragraph as strong—but not definitive—signals, and to confirm whether the narrative focus was sustained throughout the body. Optional free-text rationales were included to support auditing of ambiguous cases.

3.2 Automatic Annotation

We then evaluated two prompting configurations of the same model family, LLaMA 3.3 70B. The zero-shot configuration, *LLaMA-Base*, received the article and the list of candidate organisations and was asked to identify protagonists. The in-context configuration, *LLaMA-ICL*, received the same instruction augmented with two manually annotated exemplars. This follows the in-context learning paradigm, where demonstration examples calibrate model decision boundaries (Brown et al., 2020; Dong et al., 2024). Both configurations annotated the same 50-document sample used in the human study. Based on the comparative results, we used the in-context model to label the wider Finer-ORD corpus, including brief one-sentence justifications to enable efficient post-hoc auditing.

3.3 Agreement Study

Agreement was measured at the entity level using (i) average Jaccard similarity (mean per-document intersection-over-union), (ii) overall entity agreement (fraction of matching protagonist vs. non-protagonist labels), and (iii) Cohen’s κ to correct for chance agreement under label imbalance (Cohen, 1960). Table 2 shows representative human–human, human–LLM, and model–model comparisons.

Table 2: Pairwise agreement on the 50-document sample (411 total candidate entities).

Pair	Jaccard	Overall	κ
A1 – A2	0.371	0.786	0.406
A1 – A3	0.303	0.764	0.351
A1 – A4	0.433	0.861	0.643
A2 – A3	0.402	0.803	0.391
A2 – A4	0.331	0.779	0.378
A3 – A4	0.286	0.791	0.418
A1 – LLM-Base	0.318	0.708	0.288
A2 – LLM-Base	0.467	0.757	0.366
A3 – LLM-Base	0.334	0.691	0.201
A4 – LLM-Base	0.309	0.701	0.265
A1 – LLM-ICL	0.365	0.752	0.328
A2 – LLM-ICL	0.494	0.810	0.427
A3 – LLM-ICL	0.345	0.754	0.267
A4 – LLM-ICL	0.328	0.749	0.314
Base – ICL	0.714	0.878	0.689

Human-human comparisons show moderate to substantial agreement ($\kappa \approx 0.35$ – 0.64), confirming that protagonists are identifiable but require interpretive judgement. Overall agreement scores are inflated due to the large number of non-protagonist labels; κ therefore provides a more informative indicator of shared decision criteria.

Human-LLM comparisons show that *LLaMA-Base* over-selects protagonists, while *LLaMA-ICL* calibrates its threshold more closely to human patterns. The two model variants also show high mutual consistency ($\kappa \approx 0.69$), indicating that exemplar conditioning stabilises model behaviour.

In summary, the agreement study demonstrates that (a) PER is a reproducible discourse-level task, (b) LLMs can approximate human judgement when exemplar-calibrated, and (c) scalable corpus construction is feasible with targeted human-in-the-loop auditing.

4 Experiments

4.1 Experimental Setup

We evaluate recent open large language models on the PER task using the test splits of two datasets: CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and FiNER-ORD (Shah et al., 2024). For both datasets, we use GPT-5 to generate the “golden truth” protagonist labels, given its strong reasoning capabilities

and state-of-the-art performance among existing LLMs.

Each article provides a canonicalised list of candidate entities from which models must identify the central *protagonists*. We compare two prompting configurations: (i) a base setup with a single task instruction, and (ii) an ICL setup extending the prompt with three exemplars representing a clear single-protagonist case, a co-protagonist case, and an ambiguous case—following standard in-context learning methodology (Brown et al., 2020; Dong et al., 2024).

We test seven models spanning multiple architectures, sizes, and instruction-tuning regimes: LLaMA 4 16×17B, LLaMA 3.3 70B Instruct, LLaMA 3.1 70B, LLaMA 3.1 8B, Mistral-Nemo 12B, Gemma 3 27B, and Gemma 3 4B. All experiments use deterministic decoding (temperature = 0) to ensure reproducibility and isolate prompting effects.

Evaluation follows standard entity-level IE metrics: micro-F1, and macro-F1. Micro-F1 captures aggregate performance across all instances, while macro-F1 highlights robustness under class imbalance and document variation. Metrics are computed directly from model outputs without post-processing.

4.2 Results and Discussion

Results are presented in Table 3. For each model, we report performance under the base configuration and under ICL, and show the change relative to the base configuration.

The table reveals distinct operating regimes. Some models prioritise recall (e.g., LLaMA 3.3 Instruct and Gemma 3 27B), identifying many potential protagonists but at the cost of increased false positives. Others adopt more conservative selection policies (e.g., Mistral-Nemo 12B), improving precision but sometimes missing relevant entities.

ICL produces mixed effects. Larger models benefit most: LLaMA 4 shows substantial gains in both F1 and macro-F1 under ICL, indicating improved ability to detect protagonists across varied discourse cues. In contrast, smaller models occasionally experience performance degradation, likely due to exem-

Table 3: Protagonist detection results across datasets. Values in parentheses show the effect of using ICL relative to the non-ICL baseline, with red indicating a decrease and green an increase.

Model	ICL	CoNLL-2003		FiNER-ORD	
		Micro F1	Macro F1	Micro F1	Macro F1
Gemma-3 4B	No	0.297	0.315	0.308	0.351
	Yes	0.330 (+0.033)	0.300 (-0.015)	0.281 (-0.027)	0.270 (-0.081)
Gemma-3 27B	No	0.331	0.312	0.395	0.415
	Yes	0.351 (+0.020)	0.355 (+0.043)	0.373 (-0.022)	0.382 (-0.033)
LLaMA-3.1 8B	No	0.337	0.213	0.341	0.330
	Yes	0.320 (-0.017)	0.280 (+0.067)	0.338 (-0.003)	0.333 (+0.003)
LLaMA-3.1 70B	No	0.323	0.186	0.372	0.309
	Yes	0.394 (+0.071)	0.288 (+0.102)	0.372 (—)	0.362 (+0.053)
LLaMA-3.3 70B	No	0.401	0.317	0.408	0.433
	Yes	0.391 (-0.010)	0.363 (+0.046)	0.414 (+0.006)	0.426 (-0.007)
LLaMA-4 16x17B	No	0.225	0.105	0.434	0.371
	Yes	0.350 (+0.125)	0.207 (+0.102)	0.467 (+0.033)	0.446 (+0.075)
Mistral-Nemo 12B	No	0.375	0.270	0.429	0.423
	Yes	0.402 (+0.027)	0.254 (-0.016)	0.429 (—)	0.386 (-0.037)

plar miscalibration—ICL shifts their default inclusion threshold in ways that do not generalize across articles.

Macro-F1 deltas further show that ICL can improve consistency across documents for large models, while sometimes destabilizing smaller ones. This confirms that exemplar selection is a form of *policy shaping*: exemplars must reflect the intended balance between precision and recall.

Overall, the results show that protagonist detection is tractable with current models, but is sensitive to calibration. Larger models benefit from exemplar conditioning, while smaller models require careful exemplar selection or light post-hoc filtering. This finding motivates an annotation workflow that pairs exemplar-guided large models with selective human auditing to scale PER reliably in practice.

5 Conclusions

This paper introduced and formalised the task of *Protagonist Entity Recognition*, a discourse-level information extraction problem concerned with identifying which organisations constitute the central actors in a news narrative. In contrast to NER, which detects mentions without regard to narrative importance, protagonist detection targets *narrative centrality*: determining which entities drive the events, decisions, and consequences that

give the story its thematic focus.

We proposed an annotation framework and conducted a controlled human study across fifty news articles. The resulting agreement analysis showed that annotators converge meaningfully in identifying protagonists, indicating that the task is interpretable and reproducible. Moreover, a large model configuration demonstrated agreement with humans within the same range as inter-annotator variation, particularly when prompted with in-context exemplars. This finding supports the use of models as scalable, high-fidelity annotators.

Leveraging this insight, we automatically annotated the Finer-ORD corpus and evaluated a diverse set of recent models under matched settings. The results confirm that protagonist recognition is learnable: larger models yield robust performance, and exemplar-guided prompting provides measurable calibration benefits, especially in balancing inclusion thresholds. Smaller models, however, tend to overgeneralise protagonist roles, suggesting that their narrative reasoning remains brittle when signals are subtle or distributed across discourse segments.

Together, these contributions establish protagonist recognition as a tractable and meaningful extension of information extraction, one that moves beyond mention-level recognition toward entity-centric narrative understand-

ing. The data, evaluation protocol, and baseline results provided here create a foundation for future work on discourse-focused entity modelling, media framing analysis, organisational influence tracking, and narrative structure learning. Continued progress will likely require expanding domain coverage, refining exemplar-based calibration strategies, and integrating protagonist signals into downstream applications such as summarisation, event forecasting, and knowledge graph enrichment.

Acknowledgements

All the authors thank the annotators for their careful work and the team that assisted with prompt engineering and compute resources. The authors from Linknovate Science gratefully acknowledge support from the grant *IG408M – AXUDAS PARA O DESENVOLVEMENTO TECNOLÓXICO E A INNOVACIÓN MEDIANTE O USO DA INTELIXENCIA ARTIFICIAL (2025) – Liña A* for the project *IA para Detección de Entidades Protagonistas (PEDIA)*.



Limitations

While the results are encouraging, several limitations remain. The annotated corpus is currently limited in size and domain, focusing on financial and organisational news. Although this domain provides clear narrative structures and well-defined organisational actors, generalising to political, environmental, local, or investigative reporting may require adapting both the annotation protocol and the decision criteria for narrative centrality. The task is inherently interpretive, and absolute agreement is unlikely even among humans; richer guidelines, adjudication procedures, and annotation discussions could further stabilise the notion of a protagonist across annotators and contexts.

Our large-scale annotations rely on a single LLM configuration, which may introduce systematic biases. Although exemplar condi-

tioning improves alignment with human judgement, the choice of exemplars shapes the model’s protagonist selection policy, and different exemplar sets may produce different operating regimes. This highlights the need for systematic studies on exemplar diversity, representativeness, and robustness across models. Additionally, our evaluation assumes access to accurate candidate entity lists from upstream NER and canonicalisation; errors in these components propagate into protagonist predictions and may disproportionately affect entities with sparse or ambiguous references.

Finally, this work considers only English-language news. Narrative structures, discourse cues, and organisational naming conventions differ across languages and media ecosystems, making cross-lingual transfer non-trivial. Future work will explore multilingual protagonist detection, entity-type generalisation (e.g., persons and geopolitical entities), and the integration of protagonist signals into downstream applications such as event summarisation, influence tracing, and narrative graph construction. Beyond improving model accuracy, our aim is to position protagonist detection as a core component of broader news understanding pipelines, bridging the gap between entity recognition and narrative comprehension.

References

- Rajarshi Bhowmik, Marco Ponza, Atharva Tendle, Anant Gupta, Rebecca Jiang, Xingyu Lu, Qian Zhao, and Daniel Preotiuc-Pietro. 2024. [Leveraging contextual information for effective entity salience detection](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 395–408, Mexico City, Mexico. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Kevin Clark and Christopher D. Manning. 2015. [Entity-centric coreference resolution with model stacking](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Con-*

- ference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A survey on in-context learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- Jesse Dunietz and Daniel Gillick. 2014. [A new entity salience task with millions of training examples](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 205–209, Gothenburg, Sweden. Association for Computational Linguistics.
- Micha Elsner. 2012. [Character-based kernels for novelistic plot structure](#). In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, Avignon, France. Association for Computational Linguistics.
- Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel. 2013. [Identifying salient entities in web pages](#). In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13*, page 2375–2380, New York, NY, USA. Association for Computing Machinery.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. [Discovering diverse and salient threads in document collections](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 710–720, Jeju Island, Korea. Association for Computational Linguistics.
- Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE '97, page 10–27, Berlin, Heidelberg. Springer-Verlag.
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. [A knowledge-enhanced pretraining model for commonsense story generation](#). *Transactions of the Association for Computational Linguistics*, 8:93–108.
- Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, page 58–65, USA. Association for Computational Linguistics.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: towards zero-shot language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Agam Shah, Abhinav Gullapalli, Ruchit Vithani, Michael Galarnyk, and Sudheer Chava. 2024. [Finer-ord: Financial named entity recognition open research dataset](#). *arXiv preprint arXiv:2302.11157*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Wei Xiang and Bang Wang. 2019. [A survey of event extraction from text](#). *IEEE Access*, 7:173111–173137.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Figure 1: Initial screen showing the annotation guidelines and task overview. Annotators review examples before beginning the labelling process.

A Annotator guideline excerpt

Figure 1 presents the first screen of the annotation interface, where annotators review the task guidelines before starting. This section provides detailed instructions and practical examples illustrating how to identify the main organisational *protagonists* of each news article. The interface ensures that annotators understand the distinction between central entities—those driving the story—and peripheral mentions, promoting consistency across annotations.

Once annotators confirm their readiness, they proceed to the main evaluation interface (Figure 2). Each document displays its unique identifier, the full news article with entity mentions highlighted, and a checklist of candidate organisations extracted via NER. Annotators must select those organisations that act as the main protagonists according to the contextual cues and narrative focus of the text.

Progress indicators and validation buttons (e.g., “None are protagonists”, “Select All”, “Submit & Continue”) guide annotators through the process and ensure that each decision is explicitly reviewed before submission. This design minimises omission errors and supports a transparent, auditable labelling workflow.

Figure 2: Main annotation interface displaying a news article, highlighted entity mentions, and candidate organisations for protagonist selection.