



Generating knowledge graphs from news articles via multi-agent claim extraction and attribution

Massimiliano Fadda^a, Enrico Motta^b, Francesco Osborne^{b,c}, Diego Reforgiato Recupero^{a,*}, Angelo Salatino^b

^a Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, 09124, Italy

^b Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK

^c Department of Business and Law, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milan, 20126, Italy

ARTICLE INFO

Dataset link: <https://github.com/massi-fa/SEER>

Keywords:

Claim extraction
Multi-agent systems
Entity linking
Knowledge graph construction
Large language models
Computational journalism

ABSTRACT

News sources play a fundamental role in democratic societies by shaping how citizens access information about facts, events, and competing viewpoints. It is therefore essential to study how different outlets in different countries fulfil this role. A central aspect of such analyses is the examination of claims expressed by specific voices in the media. However, the large-scale identification, attribution, and extraction of claims from news content remain open challenges, due to the diverse ways in which claims and agents can be expressed, and the distinctive characteristics of journalistic writing. In this paper, we address this challenge by introducing SEER, a comprehensive framework for constructing large-scale knowledge graphs from news articles. To ensure high-fidelity extraction, SEER incorporates a novel Multi-Agent Ensemble module specifically designed to handle the complexity of claim extraction and attribution better than single-model approaches. The resulting knowledge graph is based on an extension of the News Classification Ontology (NCO), which captures the key characteristics of claims in news media. We also present a prototype implementation of the proposed architecture and evaluate it on a corpus of articles drawn from four domains: climate change, UK immigration, the war in Ukraine, and US politics. We perform an evaluation on manually annotated data, showing that SEER's multi-agent extraction module achieves strong performance (78.22% Macro F1), outperforming ten alternative single-model LLM-based agents on the claim extraction task. We conducted an ablation study and identified recurring error types that highlight remaining challenges in journalistic claim extraction.

1. Introduction

News sources play a fundamental role in democratic societies, as they shape how citizens access information about facts, events, and competing viewpoints. Through processes of selection, framing, and emphasis, the media influence which voices are amplified and which perspectives remain marginal [1]. When certain positions consistently dominate over others, this can substantially affect public opinion and perceptions of reality, contributing to the formation of media bubbles [2]. Such imbalances have been shown to undermine informed public debate and can have serious social and political consequences, especially when algorithmic profiling and targeted content amplify certain viewpoints and suppress balanced information [3]. For this reason, it is essential to study how individual outlets and the overall news landscape in a particular country represent viewpoint diversity, and to what extent their coverage exhibits partiality or imbalance.

Despite its importance, assessing viewpoint diversity remains a major challenge. The contemporary online news environment is characterised by an unprecedented volume of information [4], which contributes to information overload and facilitates the spread of disinformation [5]. Although automatic content analysis methods have been proposed to address this scale, they still struggle to provide comprehensive and reliable representations of the media landscape [6,7]. This has created a persistent gap between the analytical needs of media scholars, data scientists, and citizens who wish to explore these phenomena in depth, and the current capabilities of computational techniques for news analysis [8,9].

A central aspect of news analysis is the examination of claims expressed by specific voices in the media. In the context of this work, we formally define a claim as an explicitly attributed statement, opinion, or

* Corresponding author.

E-mail address: diego.reforgiato@unica.it (D. Reforgiato Recupero).

assertion made by a specific agent (either an individual or an organisation) [10]. Unlike objective descriptions of physical events or raw statistical data, a claim inherently reflects a communicative act intended to convey a specific stance, perspective, or narrative to the public. Identifying the agent, whether an individual or an organisation, responsible for a given statement, and mapping the associated claims to political or social viewpoints, is a crucial step in performing a comprehensive analysis of the range of opinions selected by news outlets. In recent years, initial work has addressed the attribution of viewpoints to sets of claims, typically using clustering techniques or LLM-based summarisation to infer underlying perspectives [11]. However, considerably less attention has been paid to the preceding and essential step that enables such approaches to operate at scale, namely the large-scale identification, attribution, and extraction of claims from news content. This task is inherently challenging due to both the ambiguity of natural language and the distinctive characteristics of journalistic writing. Claims can be expressed in multiple forms, including direct speech, indirect speech, and hybrid constructions in which the journalist's narration is interwoven with the reported words of the agent. Moreover, accurately identifying the responsible agent and linking this agent to a consistent identifier that enables the retrieval of additional contextual information is often non-trivial and highly dependent on the surrounding discourse.

In this paper, we address this challenge by introducing SEER (Semantic Extraction and Enrichment Reasoning), a comprehensive framework for constructing large-scale Knowledge Graphs from news articles. At the core of SEER lies a novel Multi-Agent Ensemble architecture specifically designed to handle the complexity of claim extraction and attribution better than single-model approaches.

Instead of relying on a single model, SEER orchestrates multiple distinct LLMs that operate in parallel as independent extraction agents. Their individual outputs are then synthesised and verified by a dedicated aggregator agent. This collaborative approach mimics a human editorial review process, harnessing the diversity of different models to mitigate hallucinations and resolve the ambiguities inherent in complex journalistic attribution.

We also present a prototype implementation of the proposed architecture and evaluate it on a corpus of articles drawn from four domains: climate change, UK immigration, the war in Ukraine, and US politics.

The architecture is organised into four main stages. First, it retrieves articles either by identifier or in response to a query. A typical analysis targets all articles within a specific topic and time span. Second, it employs an LLM-powered multi-agent ensemble to extract claims and relevant agents. This framework relies on multiple annotator agents, whose outputs are subsequently integrated and verified by an aggregator agent. Third, the system performs entity linking by mapping the extracted agents to Wikidata identifiers where possible. Finally, the resulting claim representations are used to populate a KG. This adopts, as its ontological schema, an extension of the News Classification Ontology (NCO) [10], detailed in Section 3.5.1, which we developed to formally represent the structure of utterances, agent roles, and attribution relations in news media. The resulting KG supports a wide range of analyses of claims, and voices in the media. In particular, we hypothesise that a multi-agent ensemble architecture implemented in SEER achieves superior performance in claim extraction compared to single LLM-based agents, owing to its ability to mitigate model-specific limitations such as training-data biases and hallucinations. To test this hypothesis and evaluate our full architecture, we compared our system with state-of-the-art solutions on the two core tasks of claim extraction and entity linking over the actors to whom claims are attributed. For this purpose, we constructed two gold standards by manually annotating real news articles across the four domains of interest.

The evaluation confirms our intuition, showing that SEER, leveraging its multi-agent ensemble configuration, achieves excellent performance (78.22% Macro F1), outperforming ten alternative LLM-based agents on the claim extraction task. To ensure a rigorous comparison,

these standalone agents include a wide array of LLMs, spanning diverse training provenances (including both Western and Asian developed models), and a range of access paradigms from closed-source proprietary systems to high-performance open-weight models. Moreover, SEER appears to be highly robust across all four domains and across different types of utterances, including direct, indirect, and hybrid forms, whereas individual models are more susceptible to contextual variations. The entity linking mechanism implemented in SEER also surpasses a range of alternative models, as well as a baseline based on direct lookup in Wikidata.

We conducted an in-depth error analysis of SEER, which identified recurring error types that highlight remaining challenges in journalistic claim extraction and point to directions for future work. Notably, most of the observed errors are not related to factual inaccuracies or hallucinations, but instead concern more subtle aspects of text interpretation and the generation of claims that do not fully capture the original meaning. These include failures to resolve anaphora and nested attribution, difficulties in extracting claims embedded in complex meta-commentary, and the misinterpretation of actions or opinion polls as claims. Finally, we analysed the framework's scalability and cost-performance trade-offs, showing that, while the full multi-agent pipeline is better suited to applications in which extraction robustness takes precedence over cost, a single high-performing model offers a competitive and cost-effective option for large-scale or budget-constrained analyses.

To summarise, this paper makes the following main contributions:

- We propose a novel multi-agent ensemble architecture that identifies claims in news articles, attributes each claim to a specific agent (either a person or an organisation), links these agents to their corresponding Wikidata entities, and uses the resulting information to construct a large-scale KG of the extracted claims.
- We introduce and release two manually annotated datasets that provide gold-standard annotations for the two core tasks addressed in this work, namely claim extraction and agent entity linking.
- We evaluate the proposed architecture against several alternative approaches, showing that the multi-agent ensemble consistently outperforms single-model baselines in terms of accuracy and robustness across different domains and types of utterances. We also present a qualitative analysis that categorises and discusses the main sources of error.
- We release the complete codebase of the proposed framework, named **SEER**, together with the associated datasets and the full set of system prompts used for the agents, to support reproducibility and further research.¹

The remainder of this paper is organised as follows. Section 2 provides a review of the state of the art in claim attribution, entity linking, the application of multi-agent systems to information extraction, and the semantic representation of news items. Section 3 describes the proposed architecture, detailing the end-to-end pipeline from article ingestion and multi-agent extraction to semantic enrichment and KG construction. Section 4 presents the experimental setup, introduces two novel gold-standard datasets, and provides a comprehensive analysis of the results and their implications. Finally, Section 5 summarises the main findings and outlines directions for future research.

2. Related work

This section contextualises our research within the broader landscape of automated news analysis and information extraction. Specifically, we analyse the literature according to four axes: (i) the evolution

¹ SEER Codebase, Datasets, and Prompts: <https://github.com/massi-fa/SEER>.

of claim extraction and attribution, (ii) entity linking within dynamic journalistic environments, (iii) the application of multi-agent systems, and (iv) the semantic representation of news items.

2.1. Claim extraction and attribution

The task of claim attribution, semantically binding an utterance to its source, has evolved from rigid syntactic dependency parsing to flexible generative modelling. Historically, this problem was framed as a relation extraction task, where the objective was to identify explicit links between a Source (agent), a Cue (reporting verb), and the Content. Early methodologies relied heavily on rule-based systems. Pareti et al. [12] established a foundational approach for news, demonstrating that while direct quotations can be extracted with high precision using syntactic cues, indirect and mixed quotations, which constitute nearly half of reported speech, require supervised learning to resolve complex dependencies. Similarly, O’Keefe et al. [13] proposed a sequence labelling approach that treats attribution as a decoding task, achieving high accuracy on news corpora like the Sydney Morning Herald but struggling with the indirect dialogues found in literature [14].

The advent of deep learning shifted the paradigm towards neural sequence tagging. Models based on BiLSTM-CRF [15] and BERT [16] significantly improved recall by learning contextual representations rather than relying on hard-coded patterns. Specifically, Ni et al. [15] introduced novel tagging schemes to handle overlapping relations, while Chen et al. [16] utilised candidate-specific segments to identify speakers in dense narratives. However, these discriminative approaches remain inherently limited by their extractive nature. They identify contiguous text spans but lack the capability to normalise complex, discontinuous paraphrases into structured statements. Consequently, even recent benchmarks like DirectQuote [17] largely exclude indirect speech from their scope to ensure semantic clarity, leaving the processing of complex journalistic attribution an open challenge.

The emergence of LLMs has enabled the extraction of structured claims via zero-shot or few-shot prompting, overcoming the rigidity of span-based methods. Recent state-of-the-art approaches focus on decomposing complex text into “atomic facts”. Claimify [18] introduces a framework that explicitly handles ambiguity and decontextualisation, ensuring that extracted claims are verifiable independently of their source text. Similarly, AFEV [19] proposes an iterative extraction mechanism where the model refines its understanding of a claim through a feedback loop of partial verifications.

However, the generative flexibility of LLMs introduces a critical trade-off: stochastic instability and hallucination. To mitigate this, methods like SelfCheckGPT [20] utilise stochastic sampling to detect inconsistencies in black-box models, while AGSER [21] analyses internal attention weights to perform zero-shot hallucination detection. While effective for checking factual correctness, i.e., verifying if an event occurred, these methods primarily target the content of the statement rather than the relational accuracy of the attribution. Consequently, the specific problem of misattribution, where a factual claim is assigned to the wrong speaker in a multi-agent context, remains an open challenge, particularly for black-box architectures where internal confidence scores are inaccessible.

2.2. Entity linking in news

The entity linking (EL) task, which grounds textual mentions to a knowledge base, has historically been framed as a ‘retrieve-and-rank’ classification problem (i.e., a discriminative approach). However, the advent of LLMs has fundamentally reshaped this landscape, moving the field towards generative architectures and, more recently, reasoning-based disambiguation. Despite these advancements, the application of EL to the news domain remains fraught with specific challenges regarding noise, latency, and interpretability.

A significant departure from traditional discriminative methods, which typically treat EL as a multi-class classification or ranking task over a predefined candidate space, was marked by the introduction of Autoregressive Entity Retrieval (GENRE) by Cao et al. [22]. This approach reformulates EL as a sequence-to-sequence task, where the model autoregressively generates the canonical entity name token-by-token, conditioned on the context. While this paradigm theoretically enables zero-shot linking by capturing fine-grained interactions via cross-encoding, it introduces substantial computational overhead due to the constrained beam search required during inference. To address the latency issues inherent in autoregressive models, Ayoola et al. [23] proposed ReFinED, a more efficient architecture, by performing mention detection, fine-grained entity typing, and entity disambiguation in a single forward pass. This allows it to achieve inference speeds significantly faster than previous zero-shot baselines. However, its reliance on entity descriptions creates a dependency on the quality of metadata in the Knowledge Base, which may not always be available for emerging entities in dynamic news cycles.

While models trained on Wikipedia perform well on encyclopedic text, they often degrade on noisier real-world data. Eshel et al. [24] highlight that web and news texts often lack the coherence of Wikipedia. By introducing the WikilinksNED dataset, they demonstrated that models relying solely on local context struggle when the text contains lexical overlaps or lacks explicit disambiguating cues. Similarly, Mueller and Durrett [25] argued that standard attention mechanisms often fail to identify indicative surface strings in short, noisy texts, necessitating models that can robustly handle unstructured environments.

Recent work has shifted towards leveraging the reasoning capabilities of LLMs to resolve these ambiguities. Ding et al. [26] introduce ChatEL, a multi-step prompting framework where an LLM first generates an auxiliary context and then performs disambiguation as a multiple-choice selection task. This method demonstrates that LLMs can outperform supervised baselines without fine-tuning. Building on this, Hou et al. [27] propose DeepEL, which integrates LLMs into every stage of the pipeline, including a novel self-validation mechanism. In this approach, the model uses global contextual information to cross-validate entity predictions within a sentence, correcting errors where isolated disambiguation fails.

While LLM-based reasoning yields high accuracy, the computational cost of processing every mention with a large model is prohibitive for real-time applications. To mitigate this, Li et al. [28] introduce ARTER (Adaptive Routing and Targeted Reasoning). This framework employs a lightweight router to determine the complexity of mention classification. Easy mentions are handled by a fast, traditional model, while only the complex, ambiguous cases are offloaded to an LLM for targeted reasoning. This hybrid approach attempts to balance the high precision of generative reasoning with the scalability required for large-scale text processing.

2.3. LLMs and multi-agent systems

The paradigm of automated reasoning has recently shifted from optimising single “monolithic” LLMs towards the orchestration of Multi-Agent Systems (MAS). This transition is driven by the empirical observation that a single model, regardless of its parameter size or prompting strategy, faces intrinsic limitations in self-verification and bias correction.

While techniques such as Chain-of-Thought (CoT) and self-reflection were initially proposed to enhance model reliability, recent literature highlights their inadequacy for complex disambiguation tasks. Huang et al. [29] demonstrated that LLMs struggle to correct their own responses without external feedback, often degrading performance when forced to self-edit due to intrinsic biases. This phenomenon was further analysed by Zhang et al. [30], who identified that intrinsic self-correction often causes models to waver in their responses due to

prompt bias, rather than actual error detection. Consequently, in tasks requiring high precision, single-agent architectures are prone to reinforcing initial hallucinations rather than correcting them through internal reflection.

To overcome these single-agent limitations, the ‘Society of Mind’ approach proposes leveraging diverse agents to improve factuality. Du et al. [31] established that the multi-agent debate consistently outperforms the self-consistency of the single-agent by filtering out hallucinations through logical cross-examination. Similarly, Liang et al. [32] showed that the interaction between agents prevents the ‘degeneration-of-thought’ problem, fostering more robust solutions. However, the mechanism of aggregation remains a subject of active research. Kaesberg et al. [33] systematically compared decision protocols, finding that while consensus excels in knowledge-intensive tasks, voting mechanisms are often more robust for reasoning. Furthermore, Liu et al. [34] emphasised that the benefits of collaboration are maximised not merely by adding agents, but by enforcing diversity in their reasoning strategies. Despite these advances, most of these frameworks, such as MAD-Fact by Ning et al. [35], focus primarily on logical puzzles or long-form factuality verification (QA), leaving the challenges of granular extraction from unstructured, linguistically ambiguous text largely unexplored.

The application of MAS to Information Extraction has evolved from simple conversational interfaces to specialised role-playing architectures. Early works like ChatIE by Wei et al. [36] treated extraction as a multi-turn QA problem but lacked mechanisms to resolve conflicting outputs. More recent frameworks have introduced role specialisation. For instance, Lu et al. [37] propose an architecture where agents specialise in different sub-tasks (e.g., entity typing vs. relation extraction) to resolve inconsistencies. Conversely, Wang et al. [38] focus on Zero-Shot Named Entity Recognition (NER) by employing agents specialised in different stages of the inference process (such as feature extraction and demonstration discrimination) to improve context understanding without human annotation. Current state-of-the-art systems typically frame conflict resolution as a classification correction task. They generally lack robust aggregation mechanisms designed to handle semantic ambiguity, where the challenge is not merely correcting a label, but synthesising a coherent attribution from implicit or rhetorical linguistic cues.

In contrast, SEER employs multiple independent extraction agents operating on the same unstructured input, whose outputs are reconciled by a dedicated aggregation agent explicitly designed to resolve semantic ambiguity in attribution rather than simple label disagreement.

2.4. Semantic representation of news

The semantic modelling of news content has evolved significantly, transitioning from document-centric metadata schemas to granular KGs. Current research can be categorised into three primary streams: event-centric modelling, schemas for fact-checking, and advanced ontologies for news epistemology.

The dominant paradigm in news representation focuses on extracting and structuring the “5Ws” (Who, What, When, Where, Why) of reported events. The foundational standard in this domain is the Simple Event Model (SEM), proposed by van Hage et al. [39], which provides core classes to model actors, places, and timestamps. This event-centric approach has been successfully scaled to large, multilingual environments. For instance, Rospocher et al. [40] demonstrated the extraction of Event-Centric KGs from massive streams of unstructured news, while Gottschalk and Demidova [41] aggregated heterogeneous sources to build EventKG.

Despite these advancements, recent literature highlights promising structural extensions to event representations. Building on standard event graphs, Plötzky and Balke [11] enrich event-centric knowledge graphs with narrative-aware features that accommodate conflicting roles (e.g., “freedom fighter” vs. “terrorist”) and divergent perspectives

on the same occurrence. Consequently, their approach preserves nuances of reporting, such as subjective attributions and complex framing, within the graph structure itself, rather than relegating them to unstructured metadata or losing them during aggregation. Angioni et al. [42] developed an automatic pipeline that extracts entity-relation triples from large news corpora and organises them into a knowledge graph for analysing the evolution of ESG discourse, although it does not explicitly model claim attribution or viewpoints.

With the rise of automated misinformation detection, a parallel stream focused on modelling the veracity of claims. The industry standard is currently represented by Schema.org’s ClaimReview. Tchechmedjiev et al. [43] leveraged this standard to build ClaimsKG, a large-scale graph of verified assertions. While pivotal for information retrieval, these schemas inherently suffer from a flat semantic structure. As highlighted by Arslan et al. [44] and in the survey by Guo et al. [45], existing models typically represent a claim as a monolithic text string linked to a truth value. They lack the expressivity to decompose the claim into its constituent parts (e.g., the claimer, the topic, and the stance). Arslan et al. address this by proposing the use of Semantic Frames (based on FrameNet) to capture the internal semantics of claims, yet this level of granularity remains rare in large-scale graphs.

To address the limitations of event-centric and flat models, recent theoretical work has proposed more expressive ontologies capable of capturing the epistemology of news. Early efforts like the BBC Storyline Ontology² introduced the distinction between an event and the editorial ‘storyline’. More recently, Motta et al. [10] introduced the News Classification Ontology (NCO) to formally distinguish between entities, events, situations, categorical topics, and viewpoints, offering a granular framework for analysing journalistic content. Our work builds on this line of research by operationalising and extending NCO with claim-centric constructs that support automated population from noisy journalistic text. Similarly, Motta et al. [46] proposed ontologies to identify news angles, formalising how journalists select and present specific aspects of an event.

Knowledge graph generation from unstructured text offers substantial potential but remains technically challenging. Domain-specific pipelines have been developed in several areas, including scientific literature [47,48], tourism [49], and news [42]. More recent research explores LLM-based approaches to KG construction [50,51], as well as hybrid workflows in which automatically extracted triples are validated and refined through human oversight [52]. Beyond the news domain, recent work has demonstrated the effectiveness of combining linguistic analysis with structured knowledge extraction from narrative texts. For example, Shao et al. [53] employ a pipeline integrating tokenisation, part-of-speech tagging, dependency parsing, and semantic role labelling to transform unstructured traffic-crash narratives into structured event representations and behaviour-cause relations. Their work shows how rich semantic structures can be automatically derived from noisy textual reports and subsequently exploited for downstream analytics and knowledge discovery. However, the extracted representations remain largely domain-specific and are not grounded in expressive ontologies capable of modelling viewpoints, attribution, or epistemic distinctions, which are central requirements in journalistic knowledge graphs.

A significant operational gap persists between sophisticated ontological schemas in the news domain and their deployment at scale. While LLMs offer new capabilities for information extraction, recent roadmaps by Pan et al. [54] and Bian [55] caution that robust automated pipelines capable of systematically populating such complex schemas while maintaining high semantic fidelity and structural consistency are still lacking. As noted by Bian et al. LLMs struggle to adhere to rigid, complex schemas without extensive supervision or schema-constrained decoding. When attempting to instantiate nested epistemological structures, such as distinguishing a ‘Situation’ from an

² BBC Storyline Ontology — <https://www.bbc.co.uk/ontologies/storyline>.

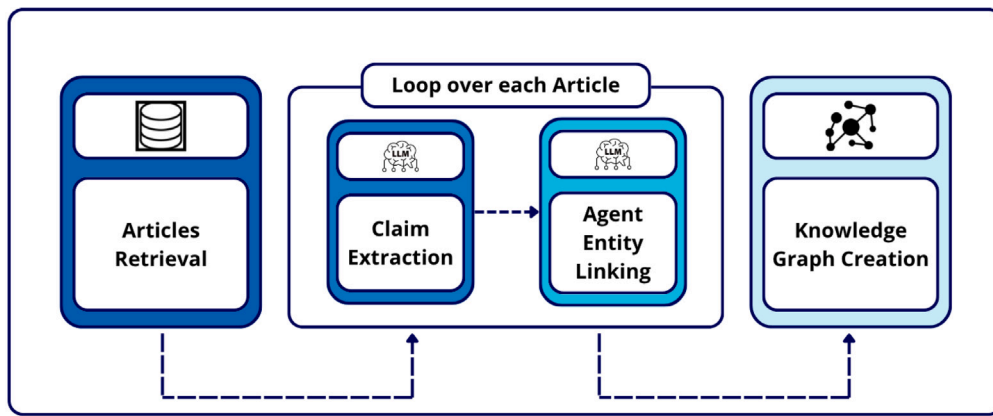


Fig. 1. The four-stage architecture of SEER. The process flows from left to right, transforming raw articles into a structured KG. The modules for claim extraction and agent entity linking leverage LLMs as their core processing engine.

‘Event’ as proposed by Motta et al. [10] from raw text, standard LLM approaches often result in hallucinations or structural inconsistencies, highlighting the need for neuro-symbolic integration.

Overall, existing approaches either focus on extraction accuracy without robust attribution guarantees, on multi-agent reasoning without grounding into structured representations, or on expressive ontologies without scalable population pipelines. To the best of our knowledge, no prior work integrates multi-agent claim extraction, attribution robustness, and semantic grounding into a KG within a unified architecture.

3. Methodology

SEER, our proposed framework, is implemented as a modular, end-to-end pipeline that systematically processes news articles, extracts claims made by individuals and organisations, and transforms them into a KG. The KG adopts, as its ontological schema, an extension of NCO [10] designed to capture the key characteristics of claims in news media. The prototype execution is orchestrated as a stateful graph using LangGraph,³ enabling robust execution, traceability of intermediate states, and transparent control over the processing flow.

The operational flow of the system is described in the following subsections. We start with a high-level architectural overview in Section 3.1. Section 3.2 introduces the semantic retrieval mechanism used to ingest relevant content. The core engine for claim extraction, which is based on a Multi-Agent Ensemble, is presented in Section 3.3. Section 3.4 details the workflow used to link persons and organisations to their corresponding representations in Wikidata. Finally, Section 3.5 explains the ontological modelling approach and the automated instantiation of the resulting KG.

3.1. System architecture overview

The overview of the system architecture is shown in Fig. 1 and includes the following stages.

1. **Articles Retrieval:** This initial stage operates at a batch level. Its function is to retrieve a target collection of articles that will be analysed by the subsequent stages. This module interfaces with a vector database for semantic search, enabling retrieval either by semantic similarity to a user’s query or by a provided list of unique article identifiers. The output of this stage is a collection of raw articles.

2. **Claim Extraction:** This stage initiates the core processing loop, employing a multi-agent ensemble of LLMs for robust analysis. For each article, multiple extractor LLMs operate in parallel to identify potential claims. An aggregator LLM then synthesises these outputs into a single, high-fidelity list. Each extracted claim is structured with key attributes, including the agent’s identity (typically the name) and type (person or organisation), the utterance (the textual content of the declaration) and its type (direct, indirect, or partially direct), and the source text for full traceability.
3. **Agent Entity Linking:** This module takes the identified agents as input and executes an LLM-powered strategy to disambiguate them by linking each agent to its canonical entity in Wikidata. This process enriches the claims in two ways. First, it assigns a stable and unique identifier to each agent. Second, it classifies each agent with a specific role, such as *Politician* or *GovernmentOrganization*, according to our predefined taxonomies for people and organisations.
4. **Knowledge Graph Generation:** Once the processing loop over all articles is complete, the final stage is executed again at the batch level. The system aggregates the fully enriched and disambiguated claims extracted from all processed articles. The module then materialises this aggregated information as an RDF graph, structured in accordance with the NCO ontology.

3.2. Articles retrieval

The entry point to our pipeline is a vector database for semantic search designed to manage and retrieve news articles. This component is a composite system comprising two key elements: (1) a primary data store that persists the full article content, including text and metadata, and (2) a high-performance vector index built using FAISS,⁴ a library optimised for efficient similarity search on dense vectors. During the ingestion process, each article’s textual content (title, body, or a combination thereof) is encoded into a high-dimensional vector embedding via a Sentence Transformers model.⁵ This vector is then added to the FAISS index along with its corresponding unique article identifier. This dual-component architecture enables fast, semantically informed retrieval through two distinct operational modes:

1. **Direct Retrieval by Identifier:** Direct retrieval allows the selection of a predefined list of specific articles, typically compiled by filtering documents according to predetermined characteristics

³ LangGraph — <https://docs.langchain.com/oss/python/langgraph/overview>.

⁴ FAISS — <https://github.com/facebookresearch/faiss>.

⁵ all-MiniLM-L6-v2 — <https://sbert.net/>.

such as a particular timeframe or subject classification. The module queries the primary data store to fetch the full content of the requested documents.

2. **Semantic Search by Query:** This mode operates on the corpus of articles that have been previously ingested and indexed. For exploratory analysis, users may also submit a natural language query. This query is converted into an embedding vector using the same Sentence Transformers model. The system then leverages the FAISS index to compute the cosine similarity between the query vector and all indexed article vectors. Retrieval is governed by a two-parameter mechanism. First, the system selects all articles whose similarity score exceeds a configurable threshold (τ), thereby ensuring a minimum level of semantic relevance. From this subset, the final output consists of the top- n most similar articles, where n is specified by the user.

The output of this stage is a collection of articles that is then passed to the core processing loop of the pipeline for analysis.

3.3. Claim extraction

Once an article is retrieved, SEER executes its Claim Extraction module. To ensure high-fidelity attribution, this module implements a Multi-Agent Ensemble strategy, as depicted in Fig. 2. Unlike monolithic approaches, this specific architectural design leverages the diversity of multiple models to enhance robustness and mitigate model-specific biases. As empirically demonstrated in Section 4, this ensemble configuration allows SEER to significantly outperform alternative approaches based on single-prompt LLMs.

To ensure clarity and reproducibility,⁶ we first define the precise data structures processed by this module.

Input and output specification. The extraction process accepts a News Article as input, comprising the article's title and its full textual body. The module transforms this input into a list of Structured Claims. Each claim is represented as a JSON object comprising the following six mandatory attributes:

- **Agent Name:** The canonical name of the person or organisation making the declaration, cleaned of titles to facilitate subsequent entity linking.
- **Agent Type:** A categorical classification of the agent, restricted to either *Person* or *Organization*.
- **Agent Description:** A brief, contextual description of the agent as presented within the source text (e.g., 'Italian Prime Minister').
- **Utterance:** The textual content of the declaration itself.
- **Utterance Type:** The linguistic classification of the declaration according to our taxonomy: *direct*, *indirect*, or *partially direct*.
- **Source Text:** The exact, verbatim sentence(s) from the article body serving as evidence, ensuring full traceability.

Taxonomy of utterances. To ensure consistent extraction and attribution, we formally define the three categories of utterances targeted by our system. These definitions are embedded in the extraction prompts to guide the agents' reasoning:

- **Direct Utterance:** The agent's statement is reported verbatim, typically enclosed in quotation marks. The attribution is explicit and the syntax preserves the speaker's original voice (e.g., *The fire chief stated, "This was the most devastating fire in decades."*).

- **Indirect Utterance:** The agent's statement is paraphrased by the journalist. While the semantic content is attributed to the agent, the grammatical structure (e.g., tense, pronouns) is integrated into the narrative flow (e.g., *The Minister confirmed that her department would launch a campaign next month.*).
- **Partially Direct Utterance:** Also known as mixed quotation, this category covers hybrid constructions where specific fragments of the agent's original wording are embedded within the journalist's narrative, retained within quotation marks for emphasis or accuracy (e.g., *The artist described her work as "a direct response" to the political turmoil.*).

The process is structured into two main phases:

1. **Parallel Extraction:** In the first phase, three distinct LLM-based agents, denoted as A, B, and C, operate in parallel as independent extractor agents. To ensure diversity in the extraction process and mitigate the bias inherent to any single architecture, these agents are instantiated using different underlying LLMs. Each agent is tasked with analysing the full text of the same article, guided by an identical, highly structured prompt. Specifically, this prompt instructs the model to identify every claim in the text and extract a structured object adhering strictly to the schema defined above. This prompt implements a form of Chain-of-Thought (CoT) reasoning [56] that guides each agent through a repeatable logical process by decomposing the extraction task into explicit steps. It also defines key concepts, such as the notion of a "claim" and the different statement types, and includes few-shot examples to ground the model's behaviour. In addition, a final checklist requires the model to self-verify its output against the core requirements before producing the final response. The prompt was developed through an iterative exploration of alternative designs, followed by empirical evaluation, and ultimately converged on the most effective solution.
2. **Aggregation and Synthesis:** The outputs of the three extractor agents, which are independent lists of candidate claims, are then passed to a fourth agent that acts as an aggregator. The role of the aggregator is to synthesise these parallel outputs into a single, consolidated, and validated list of claims. Guided by a specialised prompt that specifies how to merge and verify the inputs, the aggregator identifies consensus claims extracted by multiple agents. Crucially, this consensus is determined by the LLM's semantic understanding rather than rigid similarity metrics. This allows the system to correctly group claims that share the same meaning, even if the extractor agents used different wording. Specifically, the aggregator follows a consensus-based protocol. For metadata fields, such as agent name and type, it selects the value supported by a majority of agents. For the utterance text, it retains the exact string when at least two agents agree. When all agents produce distinct but valid variants, as often happens with indirect claims, the system integrates information from each of them to generate a final version. Finally, it discards any claims extracted by only one agent.

To ensure the pipeline's resilience, the prototype system incorporates robust error-handling mechanisms, including a retry system for transient API errors and a resilient parsing logic for potential LLM formatting mistakes. The final output of this stage is a single list of structured claim objects for the given article, which is then passed to the Agent Entity Linking module for enrichment.

3.4. Agent entity linking

In this step, the system enriches the information associated with each agent by linking it to Wikidata [57]. Wikidata was selected because it is the largest general-purpose open KG and therefore contains

⁶ The complete set of prompts used for the Multi-Agent Claim Extraction is available at: https://github.com/massi-fa/SEER/blob/main/Code/Prompts/claim_extraction_prompts.yaml.

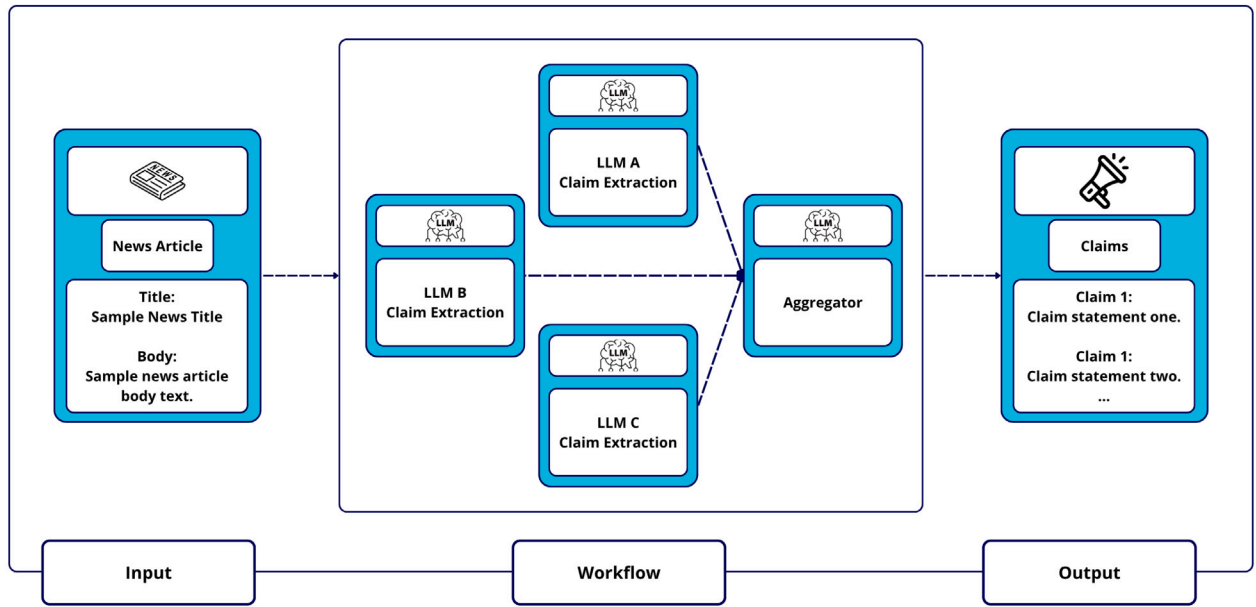


Fig. 2. SEER workflow for Claim Extraction. Three independent LLMs (A, B, and C) extract claims in parallel from the same article using an identical prompt. Their outputs are then synthesised by an Aggregator LLM to produce a single, verified list of claims as the final result.

information about a wide range of persons and organisations that are likely to be mentioned in news articles.

The objective of this stage is twofold. First, it performs entity linking by grounding each textual mention to its corresponding canonical Wikidata entity. Second, it assigns a fine-grained role to each agent through taxonomic classification.

To ensure rigorous grounding and semantic consistency,⁷ we formally define the data transformation performed in this stage.

Input and output specification. The linking module operates on a composite input consisting of the full Article Body and three specific attributes extracted from the Structured Claim in the previous stage: the Agent Name, the Agent Type (Person or Organisation), and the Agent Description.

The module returns an Enriched Agent Profile, defined as a structured object containing the following key components:

- **Wikidata Metadata:** The canonical label and unique identifier (e.g., Q12345) from the Knowledge Base. If no valid link is found, this is set to NO WIKIDATA ID (as a literal).
- **Taxonomic Classification:** A composite field containing the assigned role (from our controlled vocabularies), a confidence score (*High*, *Medium*, or *Low*) reflecting the degree of explicit evidence found in the article, and a textual explanation justifying the classification.

The inherent ambiguity of agent mentions in journalistic text makes a simple lookup strategy insufficient. We therefore designed a workflow that combines deterministic lookups, NER-based refinement, and LLM-powered disambiguation for contextual reasoning. The core entity-linking logic follows a retrieve-then-rank paradigm: candidate entities are first retrieved from Wikidata, after which a single LLM call selects the most appropriate match or returns no match if none of the candidates is suitable. This disambiguation step is essential for resolving homonymy and other forms of name ambiguity that cannot be handled through lookup alone (e.g., distinguishing between public figures with

the same name). When agent names extracted from journalistic text are noisy, an upstream NER-based fallback refines them to improve candidate retrieval. Finally, a separate LLM call performs a taxonomic classification of the agent's role, producing supplementary information independently of the entity-linking task.

The complete workflow, orchestrated as a stateful graph, is illustrated in Fig. 3. The process is designed to maximise the chances of a successful link by cascading through progressively more sophisticated techniques only when necessary.

The process unfolds through the following four strategic phases, which correspond directly to the main blocks in the architectural diagram.

3.4.1. Direct Wikidata Lookup

The entry point to the workflow is conceived as an optimistic first pass: a Direct Wikidata Lookup. The canonical agent name, as extracted from the article, is used to perform a high-speed query against the Wikidata API. This query functions as a triage mechanism, categorising each agent according to the matches returned. The process is designed to handle three distinct scenarios:

- **One Entity:** This is the ideal outcome, where the agent's name corresponds to a single, high-confidence candidate in the knowledge base. The process is then routed directly to the Validate Candidate stage for contextual verification.
- **Multiple Entities:** This result indicates ambiguity, where the agent's name matches multiple distinct entities in Wikidata (e.g., individuals with the same name but different professions). This triggers the need for a deeper analysis and the process is routed to the Disambiguate Candidates stage.
- **No Entities:** This outcome often occurs when the extracted agent name is 'noisy', it includes extraneous information common in journalistic text, such as titles ('President John Doe'), roles ('spokesperson Jane Smith'), or other descriptors. This failure to find a direct match activates the pipeline's fallback mechanism: the NER-based Refinement stage.

3.4.2. NER-based refinement

This phase acts as a fallback to improve pipeline robustness against variation in journalistic language. When a direct lookup fails, the

⁷ The complete set of prompts used for the Entity Linking and Taxonomic Classification workflow is available at: https://github.com/massi-fa/SEER/blob/main/Code/Prompts/entity_linking_prompts.yaml.

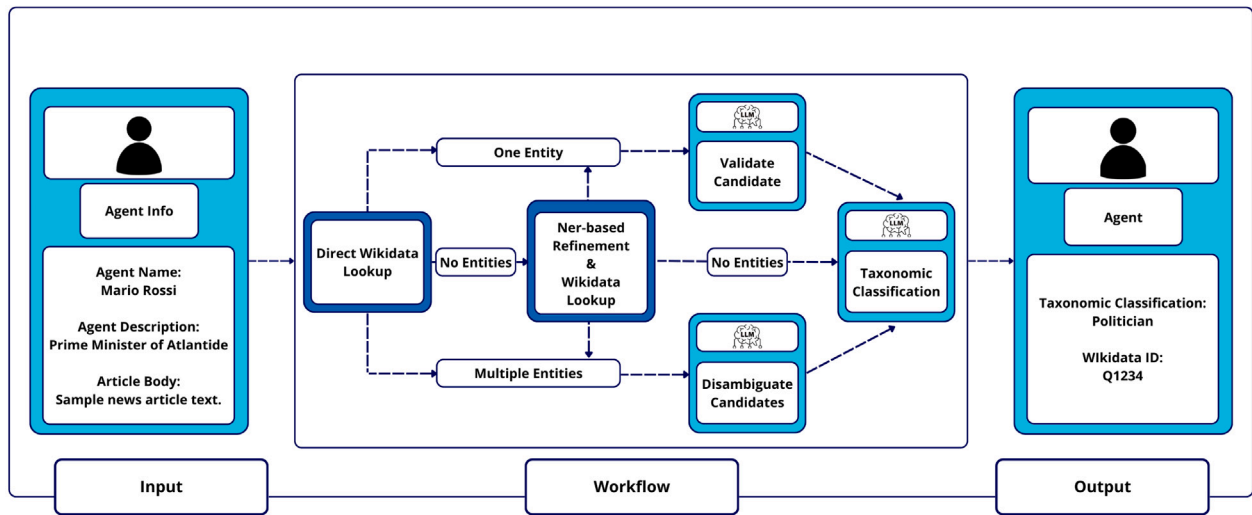


Fig. 3. SEER workflow for Entity Linking. The process follows a retrieve-then-rank paradigm: candidates are retrieved from Wikidata, and a single LLM call disambiguates among them or abstains when no valid match exists. NER-based refinement acts as a fallback to improve retrieval quality for noisy agent names. Taxonomic classification is an independent enrichment step that assigns a semantic role to every agent, regardless of linking outcome.

system infers that the entity may be present under a noisy name and transitions to the NER-based Refinement phase. In this stage, a pre-trained NER (Named Entity Recognition) model, spaCy's *en_core_web_lg*, is employed to clean the agent's name. For this purpose, the model analyses the original source text from which the claim was extracted in order to isolate the core proper noun of type PERSON or ORG (organisation). This process typically removes extraneous information that caused the initial lookup to fail, such as titles (e.g., 'Dr.', 'Senator'), roles (e.g., "spokesperson"), or affiliations (e.g., "of the White House"). The system then performs a second lookup against the Wikidata API. The result of this search is again used to determine the subsequent step in the workflow. If the query returns no entities a second time, the entity linking step terminates in order to prevent an infinite loop.

3.4.3. LLM-powered candidate resolution

This phase resolves the ambiguity of the candidates found through a process of contextual reasoning. Based on the preceding outcomes, one of two specialised LLM-powered agents is invoked. Both are provided with the full available context (agent name, description from the claim extraction stage, and the full article text) to inform their judgment. Each agent is guided by a highly structured prompt that instructs it to execute a multi-phase cognitive workflow, distinguishing between an entity's mere presence in the text and its active role in making statements.

- **Validate Candidate:** Triggered by a One Entity result, this agent's role is to act as a semantic safety net. Its primary goal is to prevent false positives by verifying if the single candidate's profile (from Wikidata) is contextually consistent with the entity mentioned in the article. It confirms the identity match even if the entity is not making a statement, but notes this distinction in its reasoning.
- **Disambiguate Candidates:** Triggered by a Multiple Entities result, this agent performs contextual adjudication. It analyses the list of potential entities, weighing their characteristics against the article's content to either select the single most appropriate Wikidata ID or conclude that none are a correct match. This process is essential for resolving ambiguity among individuals or organisations with similar names.

Both agents are engineered to return a structured JSON object containing not only the final decision (a Wikidata ID or NO WIKIDATA ID) but also a detailed reasoning field, making their decision-making process transparent and auditable. It is important to note that, while

two specialised prompt variants exist (Validate and Disambiguate) to handle different candidate scenarios, both constitute a single LLM inference step per agent. The distinction lies solely in the prompt structure, not in the number of model calls.

3.4.4. Taxonomic classification

The Taxonomic Classification node is the final, unifying enrichment stage of the workflow. It receives input from all possible preceding paths: successfully validated or disambiguated agents, as well as agents for which no entity could be linked. For every agent, an LLM is tasked with assigning a single, precise category from one of two detailed, custom-built taxonomies: NewsSpeakerType (12 roles for people) or OrganizationalSpeakerType (21 categories for organisations). The prompt instructs the model to act as an expert classifier, enforcing strict adherence to the provided controlled vocabularies. This step finalises the Enriched Agent Profile by populating the taxonomic fields with the role, confidence score, and explanation. This ensures that every agent is enriched with a meaningful semantic label, even those that could not be linked to Wikidata.

3.5. Knowledge graph generation

The final stage of the pipeline synthesises the enriched information from all processed articles into a single, cohesive KG. This module serves as the semantic backbone of our system, transforming the lists of structured claim objects into a formal, machine-readable network that ensures data interoperability and enables complex analytical queries. The construction process is guided by an ontological model that clearly separates the core entities of journalistic discourse.

3.5.1. Ontological modelling

The schema (TBox) of our KG is engineered for high semantic granularity by extending NCO,⁸ a vocabulary designed for the fine-grained representation of journalistic content. The extended version, v2, improves the previous one, v1.4.3, by (i) adding characterisations of agents (persons and organisations), (ii) explicitly introducing the notion of Utterance, and (iii) linking claims to utterances through the *hasUtterance* property.

In what follows, we give an informal description of the core classes relevant to this discussion:

⁸ News Classification Ontology — <http://data.open.ac.uk/ontology/newsclassification/v2>.

- **nco:NewsItem**: Represents the source document itself. It acts as the primary container, linking to all claims extracted from it and holding metadata such as its title and publication date.
- **nco:Claim**: Serves as the reified hub of attribution. It is defined as a subclass of **nco:Statement**, acting as a central node connecting the **nco:NewsItem** with the agent who made the statement, and the linguistic form of that statement. To ensure ontological consistency, a claim is subject to several existential and universal restrictions: it must link to exactly one **nco:NewsItem** and one **nco:Utterance**, while requiring at least one **nco:AgentComponent** and one **nco:Topic**.
- **nco:Utterance**: This node represents the actual linguistic content of the declaration. It holds the extracted statement text (**nco:hasText**) and its classification as direct, indirect, or partially direct (**nco:hasUtteranceType**).
- **nco:Agent**: This class represents the individual (**nco:Person**) or group (**nco:Organization**) that made the statement. When successfully linked, this node is connected to its canonical Wikidata entity via an **owl:sameAs** property.
- **nco:AgentRole**: This class defines the specific function or capacity of the agent (e.g., Politician, Expert) according to our custom taxonomies. These roles form a controlled vocabulary, ensuring consistency across the graph.
- **nco:AgentComponent**: This class is the key to our attribution model. It is a reified relationship that explicitly links a specific **nco:Agent** to a specific **nco:AgentRole**. This allows us to state that an agent held a particular role in the context of a specific claim, modelling scenarios where the same person might act in different capacities across different articles.

3.5.2. Instantiation and linking

The process of populating the graph with instances (ABox) is fully automated. Each entity is assigned a unique and stable Uniform Resource Identifier (URI) to prevent redundancy. Following best practices, the URIs for abstract entities like claims and utterances are generated by computing the MD5 hash of their content and context, ensuring that identical claims are represented by the same node. In contrast, URIs for concrete entities like news articles and agents are based on their human-readable identifiers.

Each enriched claim object from the pipeline is deconstructed and mapped to its corresponding RDF triples according to the ontological model described above. The final output is a single, unified graph serialised in Turtle (.ttl), chosen for its high readability, though other standard formats are also supported. An illustrative snippet of the resulting graph, demonstrating how these components are interconnected for a single claim, is provided in Fig. 4.

4. Evaluation

This section presents a comprehensive empirical validation of the proposed architecture. Section 4.1 describes the experimental protocol, including the gold standard datasets and the evaluation metrics. Sections 4.2 and 4.3 report the performance of the system on the two primary tasks: claim extraction and agent entity linking. Section 4.4 presents a detailed ablation study to rigorously validate the design choices of our multi-agent approach. Section 4.5 evaluates the complete agent-identification pipeline, comprising both the extractors and the entity-linking model, to assess error propagation between the two components. Section 4.6 discusses the computational cost and scalability of the framework. Finally, Section 4.7 demonstrates the practical utility of the generated Knowledge Graph through a downstream application focused on large-scale viewpoint and sentiment analysis.

4.1. Experimental setup

Section 4.1.1 describes the construction of the two gold standard datasets. Section 4.1.2 outlines the compared methods, including our proposal and alternative approaches. Finally, Section 4.1.3 reports the metrics used to assess performance.

4.1.1. Gold standard datasets

To ensure a robust evaluation of our architecture, we constructed two distinct gold standard datasets to assess the two core tasks of Claim Extraction and Agent Entity Linking.

Claim extraction dataset. To construct a benchmark for the claim extraction task, we first selected a dataset of 100 news articles covering four main topics: UK Immigration, War in Ukraine, US Politics, and Climate Change.

The articles were retrieved using the Aylien News API,⁹ a news intelligence platform that aggregates and indexes content from thousands of global sources. The selected corpus includes articles published between 2023 and 2025.

As illustrated in Fig. 5, the dataset provides a balanced coverage of journalistic styles, encompassing both broadsheet newspapers and agencies (e.g., Associated Press, The Guardian, BBC), which generally adopt a formal register, as well as popular press outlets (e.g., Daily Mail, The Sun), characterised by more emotive and colloquial language.

To ensure consistency and address the inherent ambiguity of journalistic text, the annotation process was governed by strict operational guidelines. Annotators were instructed to extract every atomic claim, defined as an explicitly attributed statement, opinion, or assertion. Each attributed clause was annotated as a separate claim, avoiding summarisation or aggregation of multiple claims into a single extraction.

To distinguish claims from general journalistic narrative, we established explicit boundary criteria for borderline cases:

- **Claims vs. Actions:** The description of a physical, political, or formal action (e.g., “*The Congresswoman presented a Certificate of Recognition*”) is classified as an event, not a claim. An extraction is only valid if the action is accompanied by a verbalised stance or assertion (e.g., “*The Congresswoman presented the certificate, stating it was a necessary step*”).
- **Claims vs. Statistical Reports:** Objective reporting of survey results or statistical data (e.g., “*Over half of the voters think they will change their mind*”) is not considered a claim unless it is explicitly embedded within an agent’s argumentative stance.
- **Nested Attribution:** In cases where an agent reports someone else’s statement (e.g., “*Trump said the General told him X*”), the claim is attributed to the primary agent leveraging the quote in the current discourse, rather than the historical source.

Five human annotators independently processed the articles. Inter-annotator agreement was assessed by computing pairwise Cohen’s kappa scores and averaging the results, yielding a mean value of 0.43, indicating moderate agreement. Any remaining disagreements were resolved through discussion with a sixth senior annotator. In seven cases, the news article bodies were deemed too ambiguous to support clearly identifiable claims, and those articles were excluded from the final dataset.

Notably, the moderate inter-annotator agreement primarily reflects disagreements over the boundaries of individual claims, specifically where a paraphrased or partially quoted utterance should begin and end, rather than over whether a specific claim is present. This interpretation is supported by our error analysis (Section 4.2.4): the two dominant error categories, granularity mismatch (57.3% of false positives)

⁹ Aylien News API — <https://docs.aylien.com/newsapi/v5/#getting-started>, now sunsetted.

```

@prefix nco: <http://data.open.ac.uk/ontology/newsclassification/v2#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <https://schema.org/> .

# 1. The source article
<.../NewsItem/7570061444_ukraine_war> a nco:NewsItem ;
nco:hasClaim <.../Claim/085df945303a565ae8735408fd908519> .

# 2. The central Claim node that connects everything
<.../Claim/085df945303a565ae8735408fd908519> a nco:Claim ;
nco:claimInNewsItem <.../NewsItem/7570061444_ukraine_war> ;
nco:hasAgentComponent <.../AgentComponent/Donald_Trump_Politician> ;
nco:hasUtterance <.../Utterance/2c67b2d3a52871848c56946601e293db> .

# 3. The statement (Utterance) with its text and type
<.../Utterance/2c67b2d3a52871848c56946601e293db> a nco:Utterance ;
nco:hasText "Ukraine is to blame for starting the war with Russia." ;
nco:hasUtteranceType nco:directUtterance .

# 4. The Agent Component that contextualizes the role
<.../AgentComponent/Donald_Trump_Politician> a nco:AgentComponent ;
nco:hasAgent <.../Person/Donald_Trump> ;
nco:hasAgentRole <.../AgentRole/Politician> .

# 5. The Agent, with its link to Wikidata
<.../Person/Donald_Trump> a nco:Person, schema:Person ;
rdfs:label "Donald Trump" ;
owl:sameAs <https://www.wikidata.org/entity/Q22686> .

# 6. The definition of the Role
<.../AgentRole/Politician> a nco:AgentRole ;
rdfs:label "Politician" .

```

Fig. 4. A Turtle snippet representing a single claim. It illustrates the connections between the News Item, Claim, Utterance, Agent Component, Agent (with its Wikidata link), and Role.

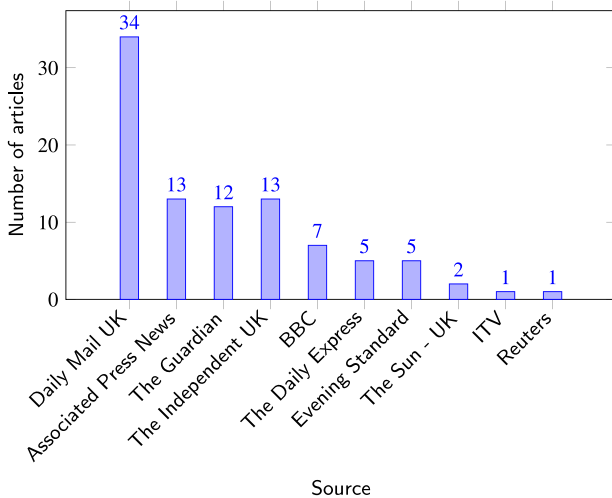


Fig. 5. Distribution of news sources.

and incomplete extraction (43.3% of false negatives), both concern utterance segmentation rather than factual inaccuracies. This effect is particularly pronounced for indirect claims, for which a journalist's paraphrase must be segmented into atomic propositions. In these cases, even small differences in the selected boundaries can alter the proposition attributed to the speaker. The observed agreement therefore reflects the intrinsic complexity of fine-grained claim segmentation, rather than fundamental disagreement about claim identification.

The resulting dataset, hereafter referred to as CED-1K (Claim Extraction Dataset),¹⁰ includes a total of 1052 claims extracted from 93 articles. It covers a variety of claim types: 559 claims are classified as direct quotations, 271 as partially direct, and 222 as indirect speech.

A detailed breakdown of the dataset's composition by topic is provided in Table 1. Notably, the table reveals significant variation in claim density, with articles on the 'War in Ukraine' containing substantially more claims on average (15.29), followed by 'US Politics' (11.46), than those on 'UK Immigration' (9.21) or 'Climate Change' (9.00). This topical diversity ensures that our evaluation is not biased towards a single narrative style and tests the system's robustness across different domains of discourse.

Each annotated claim adheres to the structured schema shown in Table 2, which defines the ground truth for our extraction task by detailing the agent, the utterance, and their context within the source article.

The articles were retrieved using comprehensive boolean queries tailored to each topic. For example, the query "(UK OR 'United Kingdom') AND (immigration OR refugee OR 'asylum seeker')" was used for the topic of UK immigration. This manually-verified dataset forms the foundation for assessing the performance of our automated system in a realistic, multi-domain context.

Entity linking dataset. To assess the performance of our multi-stage Agent Entity Linking workflow, we constructed a second dataset.¹¹ The

¹⁰ The CED-1K dataset is available for download at: <https://github.com/massi-fa/SEER/blob/main/GoldStandards/ClaimExtraction/CED-1K.xlsx>.

¹¹ The Entity Linking Gold Standard dataset is available at: <https://github.com/massi-fa/SEER/blob/main/GoldStandards/EntityLinking/EL-GS.xlsx>.

Table 1

Detailed statistics of the Claim Extraction dataset. The table details the volume of articles and claims, alongside a breakdown of linguistic types, Direct (Dir.), Indirect (Ind.), and Partially Direct (Part.), with percentages in parentheses.

Topic	Arts.	Claims	Density Mean $\pm$ SD	Claim types (Count & %)		
				Dir.	Ind.	MQ.
Climate Change	21	189	9.00 $\pm$ 6.50	117 (61.90%)	33 (17.46%)	39 (20.63%)
UK Immigration	24	221	9.21 $\pm$ 4.29	100 (45.25%)	55 (24.89%)	66 (29.86%)
War in Ukraine	24	367	15.29 $\pm$ 5.56	183 (49.86%)	80 (21.80%)	104 (28.34%)
US Politics	24	275	11.46 $\pm$ 5.30	159 (57.82%)	54 (19.64%)	62 (22.55%)
Total/Avg.	93	1052	11.31 $\pm$ 5.98	559 (53.14%)	222 (21.10%)	271 (25.76%)

Table 2

Schema for the annotated Claim Extraction gold standard. Each row in the dataset corresponds to a single claim and includes these fields.

Field name	Description
Title	The title of the source news article.
Body	The full text content of the source news article.
Agent	The name of the person or organisation making the claim.
TypeOfAgent	The classification of the agent ('person' or 'organization').
Utterance	The exact textual content of the claim being made.
TypeOfUtterance	The classification of the utterance (direct, indirect, partially direct).
SourceText	The original sentence(s) from the article containing the claim.

entity linking module is evaluated on a standalone gold standard, independently of the claim extraction pipeline. This decoupled design is a deliberate methodological choice, as it supports a clean and attributable assessment of the linking component in isolation, without confounding its performance with errors introduced in upstream extraction stages. The 90 agent instances were sampled directly from the same news corpus used for claim extraction, ensuring that the evaluation remains representative of the target domain and realistic input conditions. Specifically, the final version of the benchmark covers 78 individuals and 12 organisations, a distribution that reflects real-world news reporting.

The dataset was produced by having a human annotator manually verify the existence of a Wikidata entry for each agent. To evaluate the system's robustness against hallucination, the dataset is balanced between two distinct categories:

- **Linkable Subset (45 instances):** Cases where a verifiable Wikidata link exists. Success here requires identifying the correct canonical identifier.
- **Unlinkable Subset (45 instances):** Cases where the agent is not present in Wikidata or is too ambiguous to be linked. Success here requires the system to correctly abstain from linking (NO WIKIDATA ID prediction).

This balanced design was a deliberate choice aimed at rigorously evaluating the system's ability to avoid hallucination, a challenge in LLM-based entity linking [58]. A naturally distributed sample, in which unlinkable entities are comparatively rare, would provide insufficient negative cases to obtain a reliable estimate of hallucination rate under challenging conditions. The balanced split therefore constitutes a stress test that provides a conservative upper bound on the system's propensity to hallucinate.

Each instance was annotated according to the schema in Table 3, which provides both the ground truth for the agent's identity and the full context from the source article. Table 4 presents an overview of the different types of situations to solve for link prediction, from the case in which a single correct candidate is found to the most complex situation in which the system needs to disambiguate between different possible agents.

Table 3

Schema of the Entity Linking gold standard. Each instance includes the agent's mention, its context, and the ground truth Wikidata ID for evaluation.

Field name	Description
agent_name	The name of the agent as extracted from the claim.
agent_description	The contextual description of the agent.
agent_type	The classification of the agent ('person' or 'organization').
wikidata_id	The ground truth Wikidata ID. NO WIKIDATA ID if no correct link exists.
gs_group	A category defining the specific linking challenge scenario.
title	The title of the source news article.
body	The full text content of the source news article.

Table 4

Distribution of instances across the Linkable and Unlinkable subsets. The dataset is deliberately balanced to test the system's performance on both valid identification and noise rejection.

Challenge category	Description	Count
Linkable subset (45 instances)		
wikidata_1_result	A single, correct candidate is found via name lookup.	23
wikidata_gt1_result	The correct entity is present among multiple candidates, requiring disambiguation.	22
Unlinkable subset (45 instances)		
no_wikidata_0_result	No candidates are found for a non-linkable agent.	19
no_wikidata_1_result	A single incorrect candidate is found, which must be rejected.	10
no_wikidata_gt1_result	Multiple incorrect candidates are found, all of which must be rejected.	16

4.1.2. Tested methods

To rigorously evaluate our proposed solutions, we compared them against relevant baselines for each task.

Claim extraction. For the task of claim extraction, we evaluated several approaches that fall into two distinct categories:

- **Single-Model:** In this configuration, a single LLM is used as the sole extractor agent. It processes the article and generates a list of claims based on the structured prompt described in Section 3.3. Specifically, we tested the ten high-performing LLMs listed in Table 5 in this setting. This configuration represents the standard, non-ensemble approach to LLM-based information extraction and serves as our primary baseline.
- **SEER (Multi-Agent Ensemble Configuration):** This represents the specific instantiation of our proposed extraction methodology within the SEER framework. In this configuration, the Multi-Agent Ensemble strategy described in Section 3.3 is deployed using three distinct LLMs as parallel extractors (z-ai/glm-4.5, moonshot/kimi-k2-0905, and deepseek/deepseek-v3.1) and a fourth LLM (z-ai/glm-4.5) as the aggregator. This setup allows us to evaluate the hypothesis that a collaborative ensemble yields superior robustness compared to any single constituent model.

Table 5

Technical specifications of the 10 LLMs adopted for the evaluation. Costs are in USD per million tokens. Context window is in thousands of tokens. For MoE models, active parameters are shown in parentheses next to the total count.

Model identifier	Input (\$/M tok)	Output (\$/M tok)	Context (k tok)	Size (Params)	Reas. Cap.	Open weights
qwen/qwen3-235b-a22b-2507 [59]	0.11	0.60	262	235 B (22B)	Yes	Yes
openai/gpt-oss-20b [60]	0.04	0.15	131	21 B (3.6B)	Yes	Yes
openai/gpt-oss-120b [60]	0.04	0.40	131	117 B (5.1B)	Yes	Yes
google/gemini-2.5-flash-lite [61]	0.10	0.40	1048	N/A	Yes	No
openai/gpt-4.1-nano [62]	0.10	0.40	1047	N/A	Yes	No
meta/llama-4-maverick [63]	0.15	0.60	1048	400 B (17B)	No	Yes
deepseek/deepseek-chat-v3.1 [64]	0.20	1.00	163	671 B (37B)	Yes	Yes
z-ai/glm-4.5 [65]	0.35	1.50	131	355 B (32B)	Yes	Yes
z-ai/glm-4.6 [66]	0.40	1.75	131	355 B (32B)	Yes	Yes
moonshotai/kimi-k2-0905 [67]	0.39	1.90	128	1 T (32B)	Yes	Yes

Entity linking. For the agent entity linking task, we compared different versions of the proposed workflow against a non-contextual baseline in order to isolate the contribution of LLM-powered reasoning:

- **Direct Wikidata Lookup (Baseline):** This baseline simulates a simple entity linking approach. For each agent name, it performs a direct search using the Wikidata API and returns the first available result, if any. This baseline involves no LLM reasoning, and therefore serves to directly quantify the contribution of LLM-powered disambiguation within SEER’s workflow.
- **SEER Linking Workflow (Our Solution):** This corresponds to our full, multi-stage workflow, as described in Section 3.4. It incorporates the complete processing pipeline, including NER-based name refinement and LLM-powered candidate validation and disambiguation. The ten LLMs listed in Table 5 are used as the reasoning engines.

Adopted LLMs. The ten LLMs adopted in both experiments were chosen to represent a broad spectrum of architectures, parameter sizes, and training data. This diversity was a deliberate strategy to mitigate model-specific biases and to rigorously assess the robustness of the proposed workflows. Table 5 provides a comprehensive overview of the models employed, with column details defined as follows:

- **Model Identifier:** The unique designation used to query the model via API.
- **Input/Output Cost:** The pricing per one million tokens for prompt processing and generation, respectively.
- **Context:** The maximum context window size supported by the model (in thousands of tokens).
- **Size (Params):** The total parameter count. For Mixture-of-Experts (MoE) architectures, the value in parentheses denotes the number of active parameters.
- **Reas. Cap. (Reasoning Capabilities):** A binary indicator denoting whether the model incorporates native reasoning capabilities to enhance logical performance.
- **Open Weights:** Indicates whether the model weights are publicly accessible.

A key aspect of the model selection is the deliberate preference for models with open weights. This choice is fundamental to the scientific validity and long-term value of the study. Proprietary, closed-source models are subject to unannounced updates by their providers, which can modify their behaviour and compromise the reproducibility of results over time. In contrast, open-weight models offer a stable and transparent foundation. By using specific, versioned open models, we ensure that the experiments can be precisely replicated by the research community, a principle that is essential for scientific progress. This emphasis on openness mitigates the risks associated with the black-box nature of closed systems, avoids vendor lock-in, and enables future task-specific fine-tuning.

Table 6

Example of structural and syntactic divergence in claim extraction (Agent: Robert Jenrick).

Ground truth (Human)	System output (Multi-Agent)
Immigration is “far too high today” but targets such as David Cameron’s 2010 pledge to reduce net migration to below 100,000 were not “particularly helpful”.	Claim 1: Immigration is far too high today. Claim 2: Targets such as David Cameron’s 2010 pledge to reduce net migration to below 100,000 were not particularly helpful.

To guide the models, all prompts were engineered using a few-shot learning paradigm. This approach provides concrete examples of the desired output format and reasoning process, thereby significantly improving task adherence and the quality of the structured JSON output.

Implementation details. The workflows for the evaluation were implemented in Python 3.11 and orchestrated as stateful graphs using LangGraph. LLM interactions were managed through OpenRouter,¹² a unified API endpoint that provides access to a diverse range of models. The temperature parameter was consistently set to 0 to ensure reproducibility. With a zero temperature setting, the model output becomes deterministic, meaning that a given input always produces the same output. The NER-based Refinement stage leverages the `en_core_web_lg` model from the spaCy library.¹³

4.1.3. Evaluation metrics

We adopted two distinct evaluation procedures to assess the two primary tasks. Our evaluation protocol is designed to be robust, fair, and reflective of real-world complexities.

Claim extraction metrics. Evaluating extracted claims is challenging because success cannot be measured by a simple exact-string match. AI Agents, as well as human annotators, can indeed generate outputs that are semantically identical but syntactically different. Furthermore, the outputs may exhibit structural divergences, such as merging two distinct ground truth claims into a single sentence or splitting one complex claim into two. Table 6 illustrates these challenges with a real example. Here, the human annotator recorded a single complex claim, while the system split it into two separate, simpler sentences. Although the meaning remains identical, a traditional exact-string match would fail to recognise this as a correct extraction. This example demonstrates why our evaluation uses semantic similarity and flexible matching to account for such structural variations.

¹² OpenRouter — <https://openrouter.ai/docs/quickstart>.

¹³ `en_core_web_lg` — https://spacy.io/models/en#en_core_web_lg.

To address these challenges, we developed a two-phase matching protocol that evaluates semantic equivalence and handles structural variations.

1. **Phase 1: Agent Matching.** The first step consists of associating agents from the ground truth (GS) with those in the system results (RES). This step is essential, as claims can only be compared once their speakers have been correctly identified. The matching procedure follows a three-stage hierarchical funnel designed to maximise accuracy. First, an exact match is attempted between normalised agent names. Second, if no exact match is found, the system checks for name containment, for example, matching ‘John’ within ‘John Smith’. Finally, any remaining unmatched agents are compared semantically. Their names are encoded into vector embeddings using a Sentence Transformers model (paraphrase-MiniLM-L6-v2), and a match is confirmed if the cosine similarity exceeds a threshold of 0.7. Based on this matching logic, we frame the evaluation as an information retrieval task and define the metrics as follows:

- **True Positive (TP):** A system-extracted agent that matches a GS agent via exact, containment, or fuzzy matching.
- **False Positive (FP):** A system-extracted agent that fails to match any agent in the GS.
- **False Negative (FN):** A GS agent that is not found among the system-extracted agents.

2. **Phase 2: Claim Matching.** Once agent pairs have been established, the next step is to align their attributed claims. To accommodate the structural divergences described earlier, we employ a combinatorial matching algorithm that supports flexible many-to-many alignments. The matching process follows a greedy, largest-first search strategy. The algorithm adopts a greedy strategy, prioritising larger clusters. It iterates through combinations from size N (default: 3) down to 1, evaluating N -to- M alignments to capture fragmented or merged claims. This approach prioritises the identification of the most comprehensive and semantically informative matches. It also prevents cases in which a simple one-to-one match is confirmed prematurely, thereby avoiding the omission of a more accurate N -to- M alignment.

At each comparison level, the texts of the combined claims are encoded into a single vector representation, and a similarity matrix is computed. The algorithm selects all candidate pairs whose similarity exceeds a predefined threshold (τ), which acts as a lower bound for filtering out semantically weak alignments before the combinatorial matching step. In our experiments, we empirically set ($\tau = 0.75$), as this value provided a good balance between allowing paraphrastic matches and avoiding overly permissive alignments. To validate this choice, we conducted a sensitivity analysis by varying (τ) within a neighbourhood of this value ($\tau \in [0.65, 0.85]$) at steps of 0.05. The results showed only limited variation in the overall F1-scores and preserved the relative ranking of the evaluated methods, indicating that the conclusions of the evaluation are robust to moderate changes in the threshold.

Consistent with the information retrieval framework, the performance metrics for claim identification are defined as:

- **True Positive (TP):** A claim from the GS that is successfully aligned with one or more system-extracted claims based on the semantic threshold.
- **False Positive (FP):** A system-extracted claim that remains unmatched after the alignment process.
- **False Negative (FN):** A claim present in the GS that finds no valid counterpart in the system output.

To provide a comprehensive view of the system’s performance, we compute F1-scores using two complementary aggregation strategies:

- **Macro-Average F1:** We calculate the F1-score for each article independently and then compute the arithmetic mean across all articles. This metric treats every document as an equally important unit, ensuring that the final score reflects the system’s robustness across diverse inputs rather than being dominated by long articles with high claim density.
- **Micro-Average F1:** We sum the raw counts of TPs, FPs, and FNs across the entire corpus to calculate global Precision and Recall, from which the single global F1 is derived. This metric reflects the aggregate performance on the dataset as a whole, implicitly weighting articles according to the number of agents or claims they contain.

Entity linking metrics. For the Agent Entity Linking task, the objective is to correctly map each identified agent mention to its corresponding unique identifier in the reference Knowledge Base (Wikidata), or to correctly abstain when no valid link exists. To rigorously assess the quality of these predictions, we categorise each instance outcome as follows:

- **True Positive (TP):** The system assigns a Wikidata ID to a mention that exactly matches the ID provided in the Ground Truth. This indicates a successful retrieval. (e.g., both the system and GS map “Mario Rossi” to Q1234).
- **True Negative (TN):** The system correctly identifies that an agent cannot be linked (i.e., returns NO WIKIDATA ID) for a mention that is marked as unlinkable in the Ground Truth. This represents a correct rejection of noise.
- **False Positive (FP):** The system assigns a Wikidata ID that does not match the Ground Truth ID. This represents either a *disambiguation error* (e.g., linking “Paris” to the celebrity Q123 instead of the city Q321) or a false matching where the system incorrectly assigns an ID to an unlinkable mention (NO WIKIDATA ID).
- **False Negative (FN):** The system fails to return the correct identifier for a mention that possesses a valid ID in the Ground Truth (this includes both returning NO WIKIDATA ID or an incorrect Wikidata ID).

Based on these categories, we report performance using two complementary metrics. First, we calculate the Global Micro-F1 Score. This metric offers a holistic measure of the system’s reliability in distinguishing between known and unknown entities, treating every instance in the dataset with equal weight.

Second, to disentangle retrieval capability from noise-filtering capability, we compute Subset Micro-F1 Scores. The F1-score for the Linkable subset measures the system’s effectiveness at retrieving valid information (aggregating performance across all valid Wikidata IDs), while the F1-score for the Unlinkable subset quantifies the system’s ability to correctly abstain from forcing links for absent entities (treating NO WIKIDATA ID as the target class).

4.2. Claim extraction results

In this section, we present an empirical evaluation of the Claim Extraction task. To provide a comprehensive assessment of the system’s capabilities, the analysis is structured along four distinct dimensions. First, Section 4.2.1 reports the performance metrics. Subsequently, Section 4.2.2 examines the system’s robustness across diverse semantic domains through a fine-grained, topic-level analysis. Section 4.2.3 analyses the system’s ability to handle the syntactic complexities of direct, indirect, and partially direct attributions. Finally, Section 4.2.4 presents an error analysis that characterises the system’s failure modes and delineates the current limitations of automated claim attribution.

Table 7

Performance on the Claim Extraction task. The results for SEER reflect the performance of its Multi-Agent Ensemble module is highlighted in bold. Models are ranked by the Macro Claim F1-Score. F1-scores are reported for both agent identification and full claim extraction, using both macro and micro averaging.

Model/Configuration	Agent F1-Score		Claim F1-Score	
	Macro	Micro	Macro	Micro
SEER (Multi-Agent Ensemble)	85.74	86.36	78.22	78.88
moonshotai/kimi-k2-0905	84.97	85.26	75.73	76.97
deepseek/deepseek-chat-v3.1	84.09	84.80	75.32	76.36
z-ai/glm-4.5	80.62	81.68	72.99	74.14
z-ai/glm-4.6	79.20	80.54	72.48	73.02
openai/gpt-oss-120b	77.86	77.96	69.07	70.79
google/gemini-2.5-flash-lite	76.79	78.07	69.06	69.81
openai/gpt-oss-20b	79.23	79.82	69.04	69.21
qwen/qwen3-235b-a22b-2507	76.56	77.48	66.67	67.85
meta/llama-4-maverick	74.07	76.01	60.22	63.45
openai/gpt-4.1-nano	59.90	58.27	37.85	39.62

4.2.1. Overall performance

Table 7 reports the performance of the eleven evaluated methods on the CED-1K benchmark, covering both the prerequisite subtask of agent identification and the main task of claim extraction. The SEER framework (in its multi-agent configuration) achieves the best results, with a macro F1 score of 85.74% for agent identification and 78.22% for claim extraction. The strongest competing approaches are large open-weight models such as moonshotai/kimi-k2-0905, which attains 75.73% for claim extraction, and deepseek/deepseek-chat-v3.1, with 75.32%. In contrast, the marked performance degradation observed in smaller architectures, such as openai/gpt-4.1-nano with 37.85%, indicates the existence of a capacity threshold for the level of semantic reasoning required in attribution tasks. This result empirically supports our core premise: a consensus-driven aggregation mechanism can effectively filter the stochastic noise and hallucinations inherent in individual LLMs, yielding a higher-fidelity signal than any single model can reliably produce.

Across all configurations, F1 scores for agent identification are consistently higher than those for full claim extraction. This pattern suggests that, while identifying the speaker constitutes a comparatively bounded NER task, accurately capturing the complete semantic boundaries of their utterances, without truncation or hallucination, remains the primary challenge in automated news analysis.

To rigorously validate these performance differences, we conducted a paired bootstrap resampling test (10,000 iterations) at the article level ($N = 93$). The analysis confirms that the performance gain of SEER (Multi-Agent Ensemble) over the best single model (Kimi-K2) is statistically significant ($p = 0.022$, 95% CI: [+0.06, +5.06] percentage points). Furthermore, the ensemble significantly outperforms its own aggregator model (GLM-4.5) acting as a standalone extractor ($p < 0.001$, 95% CI: [+2.04, +8.65]), confirming that the performance gain is genuinely driven by the multi-agent synthesis rather than the capability of the aggregator alone.

4.2.2. Performance by topic

Table 8 reports the performance of the evaluated system on the final claim extraction task across the four topics in the gold standard: *Climate Change*, *UK Immigration*, *War in Ukraine*, and *US Politics*. This analysis enables an assessment of the robustness of the different methods across domains that exhibit distinct linguistic styles, rhetorical conventions, and domain-specific terminology.

A clear trend emerges in the strong cross-domain robustness of SEER, which achieves the best performance in three of the four topics (*Climate Change*, *War in Ukraine*, and *US Politics*) and ranks a close second in *UK Immigration*. In contrast, individual models exhibit notable domain sensitivity. For example, moonshotai/kimi-k2-0905 performs particularly well on *UK Immigration* (F1: 78.55%) but experiences

a substantial drop on the more linguistically complex *US Politics* dataset (F1: 69.05%). SEER's ensemble approach effectively attenuates this volatility by aggregating diverse reasoning strategies, thereby reducing model-specific biases and ensuring more reliable extraction across narrative styles.

Notably, both z-ai/glm-4.5 and openai/gpt-oss-20b exhibit an unusual performance pattern, characterised by a pronounced imbalance between precision and recall. Together, these two models secure the highest recall across all topics, with glm-4.5 leading in *Climate Change* and *UK Immigration*, and gpt-oss-20b in *War in Ukraine* and *US Politics*. However, this exhaustive extraction capability comes at the cost of significantly lower precision, resulting in only moderate overall F1 scores. A similar, albeit less severe, pattern is observed for z-ai/glm-4.6 and openai/gpt-oss-120b.

Certain topics also appear intrinsically easier for LLMs to process, reflecting differences in linguistic structure across domains. In particular, *Climate Change* consistently yields the highest scores (Ensemble F1: 83.80%), which we attribute to the formal and structured nature of scientific discourse, including explicit attribution verbs and standardised citation patterns. By contrast, *US Politics* emerges as the most challenging domain (Ensemble F1: 74.70%). Political discourse frequently involves rhetorical nuance, sarcasm, and indirect attributions, where the relationship between the speaker and the statement is often implicit, making accurate claim extraction more difficult.

4.2.3. Performance by utterance type

Table 9 reports the performance of the eleven systems across the three utterance types: direct, indirect, and partially direct. As previously discussed, indirect and partially direct speech is expected to be more difficult to extract. Indeed, several previous efforts have focused exclusively on recognising direct claims [13,14,17].

An interesting outcome of this evaluation is that the systems do not exhibit substantial performance differences across utterance types. This may be due to the strong capabilities of modern LLMs in recognising linguistic variation. As in the topic-based analysis, SEER proves robust across all utterance types, achieving the highest F1 score in each category. The performance of the other systems also remains relatively stable across the different utterance types. As before, glm-4.5, glm-4.6, gpt-oss-120b, and gpt-oss-20b exhibit an imbalance between precision and recall that is consistent across categories.

4.2.4. Error analysis

To better understand the behaviour and limitations of SEER's extraction module, we conducted a two-stage error analysis on the mismatches produced by the full-ensemble configuration on the CED-1K dataset.

In the first stage, we manually inspected a stratified random sample of 50 mismatches, comprising 25 false negatives (FNs) and 25 false positives (FPs), drawn proportionally from the four topics. This analysis was used to identify recurring failure patterns and organise them into a taxonomy of error types. In the second stage, we applied the resulting taxonomy to a broader sample of 150 FNs and 150 FPs, again drawn from the full evaluation set, using a prompted LLM to classify each mismatch according to the identified error categories. This allowed us to measure the relative frequency of each error pattern.

Analysis of false negatives. False negatives correspond to claims that are present in the gold standard but were not successfully recovered as matching extractions by the system. Our qualitative analysis identifies six recurring FN patterns: incomplete extraction, minor-agent miss, nested attribution, unresolved anaphora, meta-commentary, and other. We discuss each pattern below and report relevant examples.

Incomplete Extraction. This category includes cases in which the system correctly identifies the agent but recovers only part of the annotated utterance. The resulting output is partially aligned with the gold claim, but remains too narrow. A representative example is Roz Foyer's statement on Grangemouth, where the system preserves the opening proposition but omits the second clause specifying Unite's proposal concerning sustainable aviation fuel.

Table 8

Topic-level performance on the Claim Extraction task. Each sub-table shows the Macro Precision (P), Recall (R), and F1-Score for claims for a specific topic. The proposed SEER (Multi-Agent Ensemble) is highlighted in bold.

(a) Climate Change				(b) UK Immigration			
Model	P	R	F1	Model	P	R	F1
SEER (Multi-Agent Ensemble)	78.65	93.76	83.80	kimi-k2-0905	77.57	83.77	78.55
deepseek-v3.1	75.04	91.66	80.85	SEER (Multi-Agent Ensemble)	76.51	80.04	76.09
kimi-k2-0905	75.44	86.03	77.97	z-ai/glm-4.6	68.03	86.95	74.75
glm-4.6	66.85	93.85	75.15	z-ai/glm-4.5	66.85	89.74	74.88
gpt-oss-120b	65.38	93.60	75.39	deepseek-v3.1	75.80	74.59	72.05
glm-4.5	64.13	96.15	74.02	gemini-2.5-flash-lite	65.48	84.34	71.80
gpt-oss-20b	65.89	90.88	73.70	gpt-oss-120b	59.61	81.02	66.15
gemini-2.5-flash-lite	65.29	90.28	72.16	gpt-oss-20b	61.58	78.99	66.42
qwen3-235b-a22b-2507	62.69	89.51	68.50	llama-4-maverick	67.73	71.89	67.73
llama-4-maverick	56.07	67.64	59.28	qwen3-235b-a22b-2507	54.67	77.96	61.45
gpt-4.1-nano	38.39	47.79	39.45	openai/gpt-4.1-nano	39.37	49.12	41.63
(c) War in Ukraine				(d) US Politics			
Model	P	R	F1	Model	P	R	F1
SEER (Multi-Agent Ensemble)	80.08	81.16	78.99	SEER (Multi-Agent Ensemble)	73.79	80.79	74.70
kimi-k2-0905	82.63	77.08	77.63	deepseek-v3.1	69.26	79.99	71.82
deepseek-v3.1	79.23	78.72	77.25	kimi-k2-0905	67.36	77.00	69.05
glm-4.5	72.43	82.29	74.14	glm-4.5	62.36	85.25	69.03
qwen3-235b-a22b-2507	73.86	75.66	72.13	glm-4.6	64.89	81.78	68.90
glm-4.6	65.47	80.39	71.44	gpt-oss-20b	57.53	86.87	67.10
gemini-2.5-flash-lite	67.97	77.97	71.04	gpt-oss-120b	57.87	82.96	64.93
gpt-oss-120b	64.88	81.43	70.59	qwen3-235b-a22b-2507	60.04	78.96	64.84
gpt-oss-20b	61.58	82.96	69.54	gemini-2.5-flash-lite	56.68	75.67	61.61
llama-4-maverick	63.49	60.15	60.31	llama-4-maverick	53.53	55.58	53.44
gpt-4.1-nano	45.80	42.34	41.93	gpt-4.1-nano	29.63	31.21	28.58

Table 9

Detailed performance breakdown by Utterance type (Micro-Average). The table compares Precision (P), Recall (R), and F1-Score across direct, indirect, and partially direct claims. **SEER (Multi-Agent)** achieves the best balance, significantly improving precision over high-recall single models like GLM-4.5.

Model	Direct			Indirect			Partially Direct		
	P	R	F1	P	R	F1	P	R	F1
SEER (Multi-Agent Ensemble)	78.17	80.97	79.54	77.24	81.90	79.50	76.64	80.44	78.50
moonshotai/kimi-k2-0905	75.80	79.77	77.74	75.57	80.93	78.16	75.08	79.93	77.43
deepseek/deepseek-chat-v3.1	74.80	79.25	76.96	73.76	79.03	76.31	73.86	77.83	75.79
z-ai/glm-4.5	66.13	85.99	74.76	65.39	90.03	75.76	66.78	87.50	75.75
z-ai/glm-4.6	66.29	83.41	73.87	65.03	85.57	73.90	66.88	84.22	74.55
openai/gpt-oss-120b	61.76	84.00	71.19	62.09	86.11	72.15	61.12	84.85	71.05
openai/gpt-oss-20b	59.82	84.78	70.15	60.21	86.19	70.90	59.79	84.75	70.11
google/gemini-2.5-flash-lite	63.59	80.15	70.92	63.29	81.98	71.43	64.66	79.12	71.16
qwen/qwen3-235b-a22b-2507	62.32	77.34	69.02	62.24	78.59	69.46	63.38	76.69	69.40
meta/llama-4-maverick	68.49	59.85	63.88	68.12	60.71	64.20	67.56	60.11	63.62
openai/gpt-4.1-nano	39.16	41.72	40.40	39.04	42.75	40.81	40.04	42.00	41.00

- **Source Text:** ‘Ms Foyer insisted there was still a viable option on the table for Grangemouth, saying the union Unite had proposed to both the Scottish and UK governments that it could be used to become a production centre for sustainable aviation fuel.’
- **Gold Standard Agent:** Roz Foyer
- **Gold Standard Claim:** ‘there was still a viable option on the table for Grangemouth, the union Unite had proposed to both the Scottish and UK governments that it could be used to become a production centre for sustainable aviation fuel.’
- **System Output Agent:** Roz Foyer
- **System Output Claim:** ‘There is still a viable option on the table for Grangemouth.’

Minor-Agent Miss. This category captures cases in which a claim is present in the article, but the extractor fails to retain it because the speaker is anonymous, generic, or otherwise peripheral to the article’s main narrative. The issue is not hallucination, but salience bias: the system appears to prioritise prominent, or at least clearly named, speakers over weaker or less distinctive sources. A representative example is a statement attributed to an anonymous senior government source in an article otherwise dominated by claims from Donald Trump and a Conservative MP.

- **Source Text:** ‘A senior government source told The Telegraph: “We have always said we want to put Ukraine in the strongest position, whether that’s for going into talks or for the continuation of fighting. We still don’t know which of those positions we’re going to be in.”’
- **Gold Standard Agent:** Senior government source
- **Gold Standard Claim:** ‘We have always said we want to put Ukraine in the strongest position, whether that’s for going into talks or for the continuation of fighting.’
- **System Output Agent:** [Agent Missing]
- **System Output Claim:** [Claim Missing]

Nested Attribution. This category refers to cases in which one speaker reports, paraphrases, or endorses the words of another speaker, and the system collapses the attribution structure to the wrong level. In such cases, the extracted proposition may be semantically valid, but evaluation fails because the system attaches it to the inner source rather than the outer reporting speaker, or vice versa. A representative example is Donald Trump quoting Dan Caine in a nested reporting structure.

- **Source Text:** “We’re only hitting them from a temporary base in Syria,’ Trump said Caine told him.”

- **Gold Standard Agent:** Donald Trump
- **Gold Standard Claim:** “We’re only hitting them from a temporary base in Syria”
- **System Output Agent:** Dan Caine
- **System Output Claim:** “We’re only hitting them from a temporary base in Syria. But if you gave us permission, we could hit them from the back, from the side, from all over - from the base that you’re right on, right now, sir. They won’t know what the hell hit them.”

Unresolved Anaphora. This category covers cases in which a claim is introduced by a pronoun, a generic noun phrase, or another context-dependent referring expression. Although the system can generally handle such cases, it occasionally fails to resolve the reference to the appropriate named entity. As a result, the claim is attached to an unresolved or generic agent rather than to the specific speaker identified in the gold standard. A representative example is shown below. In this case, the system extracts the correct proposition but assigns it to the generic syntactic subject (‘Ministers’) instead of resolving the appositive reference to the specific speaker (‘Suella Braverman’).

- **Source Text:** ‘Ministers here in Britain, including Home Secretary Suella Braverman, have said the global number of migrants is directly linked to those who end up attempting to cross the Channel.’
- **Gold Standard Agent:** Suella Braverman
- **Gold Standard Claim:** ‘the global number of migrants is directly linked to those who end up attempting to cross the Channel.’
- **System Output Agent:** Ministers
- **System Output Claim:** ‘The global number of migrants is directly linked to those who end up attempting to cross the Channel.’

Meta-Commentary. This category includes cases in which the system misses an utterance because it is embedded in journalistic framing, paraphrase, or reported narration rather than presented as a straightforward attributed quotation. Such cases are particularly challenging because the extraction system must separate the claim-bearing content from the surrounding narrative scaffolding and stylised reporting verbs. We report below an example involving a summarised list of statements made by Donald Trump.

- **Source Text:** ‘It comes after the Republican slammed Ukrainian President Zelensky’s approval ratings, called for fresh Ukraine elections and appeared to suggest that Ukraine started the war with Russia following the historic peace talks in Saudi Arabia.’
- **Gold Standard Agent:** Donald Trump
- **Gold Standard Claim:** ‘Zelensky’s approval ratings are low. Ukraine needs fresh elections. And it looks like Ukraine started the war with Russia after those historic peace talks in Saudi Arabia’
- **System Output Agent:** [Agent Missing]
- **System Output Claim:** [Claim Missing]

This example clearly illustrates the system’s limitation in recovering claims that are heavily paraphrased and packaged within the journalist’s editorial narrative. Because the text relies on descriptive verbs (“slammed”, “called for”, “appeared to suggest”) rather than explicit quotation marks or standard attribution markers, the system bypasses the text entirely, treating it as background context rather than a specific agent’s extractable claim.

Other. This category was defined to allow the classifier to avoid forcing a mismatch into one of the other categories when none was appropriate. It includes borderline cases, annotation artefacts, and mismatches that do not fit cleanly into the main taxonomy.

Analysis of false positives. False positives are cases in which the system extracts a claim that does not match any claim in the gold standard. Our analysis indicates that these errors are rarely pure hallucinations. Instead, they usually arise from two main sources: disagreements over the boundaries of otherwise relevant claims, and cases in which the system mistakenly treats content as a claim even though it does not satisfy the criteria of the annotation scheme. We discuss below the categories that are most representative of the emergent error patterns.

Granularity Mismatch. This category covers cases in which the extracted text is directly supported by the article and correctly attributed, but the system extracts only part of the utterance reported in the gold standard. In particular, the system sometimes isolates specific propositions within broader utterances, thereby producing outputs that are semantically relevant but too narrow to match the annotated claim. We report below a typical example.

- **Source Text:** “Repeated failure to be transparent makes proper democratic scrutiny impossible. It’s vital we’re all given access to the information behind the claims made by ministers when they make them, and that any data quoted publicly is freely available. Public trust in politics is low. Our democracy depends on good information and right now we have a system that operates as if transparency and accountability are optional.”
- **Gold Standard Agent:** Andrew Dudfield
- **Gold Standard Claim:** “Repeated failure to be transparent makes proper democratic scrutiny impossible. It’s vital we’re all given access to the information behind the claims made by ministers when they make them, and that any data quoted publicly is freely available. Public trust in politics is low. Our democracy depends on good information and right now we have a system that operates as if transparency and accountability are optional.”
- **System Output Agent:** Andrew Dudfield
- **System Output Claim:** ‘Public trust in politics is low.’

Action-as-Claim. This category includes cases in which the system confuses a physical action, a procedural step, or a journalistic description of an event with a propositional claim. Although the extracted content remains strictly grounded in the source text, it violates the ontological boundary of what constitutes an attributable discourse act under the annotation scheme. We report an example below.

- **Source Text:** “During the committee hearing, Chairman Brian Mast (R-Fla.) played a video showcasing examples of taxpayer-funded programs, including a drag show workshop for Venezuelan migrants in Ecuador, which he cited as a clear instance of government waste.”
- **Gold Standard Agent:** [No corresponding agent in gold standard]
- **Gold Standard Claim:** [No corresponding claim in gold standard]
- **System Output Agent:** Brian Mast
- **System Output Claim:** “I played a video showcasing examples of taxpayer-funded programs, including a drag show workshop for Venezuelan migrants in Ecuador, which is a clear instance of government waste.”

In this case, the system does not hallucinate the event. However, it forces this narrative description into the schema of an extracted claim and performs a syntactic shift, converting the journalist’s third-person reporting into an artificial first-person utterance.

Poll-as-Claim. This category arises when the system treats survey findings, measurement results, or aggregate statistical observations as if they were specific claims made by an agent. Although such content is factual and may be loosely attributable to the organisation publishing it, it does not qualify as claim material under an annotation scheme centred on rhetorical or propositional discourse. We report an example below.

Table 10
Distribution of false negative error types.

FN category	Count	Share (%)
Incomplete extraction	65	43.3
Minor-agent miss	37	24.7
Other	23	15.3
Nested attribution	10	6.7
Anaphora unresolved	9	6.0
Meta-commentary	6	4.0
Total	150	100.0

Table 11
Distribution of false positive error types.

FP category	Count	Share (%)
Granularity mismatch	86	57.3
Action-as-claim	32	21.3
Other	28	18.7
Poll-as-claim	4	2.7
Total	150	100.0

- **Source Text:** “The fact-checking organisation Full Fact has obtained new freedom of information data which shows that between 1 January and 7 November 2022 only around 1% of all males arriving on small boats at Western Jet Foil claimed to be under 18 but were then later found to be over 18.”
- **Gold Standard Agent:** [No corresponding agent in gold standard]
- **Gold Standard Claim:** [No corresponding claim in gold standard]
- **System Output Agent:** Full Fact
- **System Output Claim:** “Between 1 January and 7 November 2022 only around 1% of all males arriving on small boats at Western Jet Foil claimed to be under 18 but were then later found to be over 18.”

Other. As before, we define an *other* category for cases that do not fall under any of the remaining categories.

Quantitative analysis. Tables 10 and 11 report the distribution of the 150 FN and 150 FP cases across the previously defined error categories.

For FNs, *incomplete extraction* accounts for 43.3% of cases, confirming it as the dominant failure mode. *Minor-agent miss* contributes a further 24.7%, indicating that secondary or embedded speakers remain a persistent blind spot. The remaining categories (excluding “Other”) are individually less frequent but collectively account for 16.7% of FNs, pointing to consistent difficulties with discourse-level structure.

For FPs, *granularity mismatch* is the dominant category, accounting for 57.3% of cases and reflecting a systematic tendency towards over-segmentation or over-inclusion of text. *Action-as-claim* follows at 21.3%, confirming that the boundary between reportable actions and verbalised stances is a recurring source of confusion.

In conclusion, the analysis reinforces the qualitative observation that SEER’s extraction errors are best understood as problems of segmentation, attribution, anaphora resolution, and the interpretation of reported discourse, rather than as cases of pure hallucination or fabrication by the underlying models.

4.3. Entity linking results

Table 12 reports the performance of the eleven evaluated methods on the Entity Linking task. The most notable result is the substantial performance gap between LLM-based agents and the baseline approach based on Wikidata Search. While the baseline achieves decent recall on valid entities, it fails significantly on the Unlinkable subset (F1-Score: 54.84%). This finding highlights the limitations of simple lookup strategies and underscores the necessity of contextual reasoning, which allows our proposed workflow to correctly identify when not to link an agent.

Table 12
Entity Linking performance ranked by Global Micro-F1. The breakdown of F1-Scores for Linkable and Unlinkable subsets highlights the models’ robustness in handling both valid links and entity rejection.

Model	Global Micro F1	Subset Micro F1	
		Linkable	Unlinkable
openai/gpt-oss-20b	94.44	93.33	95.56
z-ai/glm-4.6	93.33	93.48	93.18
openai/gpt-oss-120b	92.22	90.32	94.25
deepseek/deepseek-chat-v3.1	90.00	89.36	90.70
moonshotai/kimi-k2-0905	90.00	89.58	90.48
z-ai/glm-4.5	90.00	88.17	91.95
qwen/qwen3-235b-a22b-2507	88.89	87.50	90.48
google/gemini-2.5-flash-lite	84.44	83.67	85.37
meta-llama/llama-4-maverick	83.33	82.98	83.72
openai/gpt-4.1-nano	74.44	75.25	73.42
<i>Baseline (Wikidata Search)</i>	62.22	66.10	54.84

The LLM-powered agents achieve consistently strong results. A leading tier is formed by openai/gpt-oss-20b and z-ai/glm-4.6. The former emerges as the top performer with a Global F1 of 94.44%, while the latter follows closely at 93.33%. Their performance is characterised by a high balance across challenge types: gpt-oss-20b, for instance, attains an F1-Score of 93.33% on Linkable agents and 95.56% on Unlinkable ones. These symmetric values indicate a robust ability to correctly identify valid entities while reliably rejecting non-linkable mentions, which is essential for a trustworthy entity linking system.

A clear performance hierarchy emerges across all evaluated models. Configurations such as openai/gpt-oss-120b, moonshotai/kimi-k2-0905, and deepseek/deepseek-chat-v3.1 also deliver strong results, with Global F1 consistently reaching or exceeding 90%. In contrast, smaller models such as openai/gpt-4.1-nano perform substantially worse, dropping to an F1 of 74.44% due to a reduced capability in handling unlinkable entities (F1-Score: 73.42%). These results provide an empirically grounded basis for selecting suitable models for high-stakes entity linking applications.

4.4. Ablation study

To validate the core design choices of SEER and assess the contribution of its main architectural components, we conducted an ablation study. The study is organised into four thematic blocks:

- Aggregator choice:** assesses the effect of the aggregation stage by keeping the extractor configuration fixed and varying only the aggregator model.
- Homogeneous ensembles:** tests whether SEER’s gains derive from heterogeneous model composition or from repeated inference by comparing SEER-Full with homogeneous ensembles in which the same extractor model is run three times.
- Ensemble size:** evaluates the marginal contribution of the third extractor by comparing SEER-Full with all possible two-extractor subsets.
- Weak-agent robustness:** examines how performance changes when one top-tier extractor is replaced with a weaker model.

The three models composing the SEER-Full ensemble are referred to throughout this section as: GLM-4.5 (z-ai/glm-4.5), Kimi-K2 (moonshotai/kimi-k2-0905), and DS-V3.1 (deepseek/deepseek-chat-v3.1). Block D additionally introduces GPT-20b (openai/gpt-oss-20b), a mid-tier model used to probe the robustness of the ensemble when one extractor is replaced by a substantially weaker agent. Full specifications for all models are reported in Table 5.

All experiments were evaluated on the full CED-1K gold standard using the same metrics as the main evaluation (F1). To ensure a fair comparison, extractor outputs were cached at temperature $T = 0$ and

Table 13

Results of the ablation study conducted on the CED-1K gold standard. Each configuration varies a single design dimension with respect to SEER-Full: the aggregator model (Block A), the degree of architectural diversity among extractors (Block B), the number of extractor agents (Block C), and the quality of the weakest extractor (Block D). Δ denotes the absolute difference in F1-Score relative to the baseline. Configurations in Block B use $T = 1.0$ to maximise stochastic variation; all others use $T = 0$.

ID	Block	Extractors	Aggregator	F1	
				Macro	Δ
P	Baseline	GLM-4.5 + Kimi-K2 + DS-V3.1	GLM-4.5	78.22	–
Block A — Aggregator sensitivity					
A2	Agg.	GLM-4.5 + Kimi-K2 + DS-V3.1	Kimi-K2	77.41	–0.81
A1	Agg.	GLM-4.5 + Kimi-K2 + DS-V3.1	DS-V3.1	76.45	–1.77
Block B — Homogeneous ensembles ($T = 1.0$)					
B1	Homo.	Kimi-K2 $\times$ 3	DS-V3.1	78.11	–0.11
B2	Homo.	DS-V3.1 $\times$ 3	Kimi-K2	76.10	–2.12
B3	Homo.	GLM-4.5 $\times$ 3	Kimi-K2	75.86	–2.36
Block C — Heterogeneous 2-Agent ensembles					
C1	2-Agt.	Kimi-K2 + DS-V3.1	GLM-4.5	76.06	–2.16
C3	2-Agt.	GLM-4.5 + DS-V3.1	Kimi-K2	74.82	–3.40
C2	2-Agt.	GLM-4.5 + Kimi-K2	DS-V3.1	73.94	–4.28
Block D — Robustness to weaker agents					
D1	Weak	Kimi-K2 + DS-V3.1 + GPT-20b	GLM-4.5	77.14	–1.08

reused across configurations that share the same extractors. Configurations in Block B use $T = 1.0$ to maximise stochastic variation across repeated runs of the same model.

The full set of configurations and their results is reported in Table 13. In the following, we discuss the results for each of the four blocks.

4.4.1. Effect of aggregator choice

Block A isolates the contribution of the aggregator by keeping the three heterogeneous extractors fixed and substituting only the aggregator model. Replacing GLM-4.5 with Kimi-K2 (A2) yields a modest decrease of -0.81% in F1 (77.41%), while substituting DS-V3.1 (A1) produces a larger drop of -1.77% (76.45%), confirming that the choice of aggregator has a measurable and consistent impact on overall performance.

What makes this result particularly informative is the identity of the best-performing aggregator. GLM-4.5 is the weakest of the three models as a standalone extractor, achieving 72.99% F1 against 75.73% for Kimi-K2 and 75.32% for DS-V3.1 (Table 7), yet it produces the strongest aggregated output when placed in the aggregator role. This dissociation between extraction and aggregation performance suggests that the two roles draw on complementary capabilities: synthesising and reconciling outputs from multiple agents appears to require a different balance of skills than independently identifying claims from raw text.

The ordered degradation observed as the aggregator is replaced by progressively weaker models further supports the conclusion that GLM-4.5's advantage in this role reflects a genuine integration effect. Since GLM-4.5 also participates as one of the three extractors, its outputs are already represented in the input to the aggregator; if it were simply amplifying its own prior extractions, swapping it out would have little effect. The fact that it does not confirms that the aggregated output is the product of cross-model synthesis rather than single-model dominance.

4.4.2. Homogeneous ensembles

Block B tests whether the performance advantage of SEER-Full over single-model baselines is driven by the architectural diversity of the extractor models, or whether similar results could be obtained by simply running the same model multiple times. To isolate this factor, we instantiate three homogeneous ensembles in which the same model is used three times as an extractor at temperature $T = 1.0$, maximising

stochastic variation across runs and providing the strongest possible repetition-based counterpart to the heterogeneous configuration.

While configurations B2 and B3 fall significantly short of the baseline, configuration B1 (Kimi-K2 $\times$ 3) achieves 78.11%, performing almost identically to SEER-Full (78.22%). To formally validate this observation, we conducted a paired bootstrap resampling test (10,000 iterations), which confirms that this 0.11% difference is indeed not statistically significant ($p = 0.470$, 95% CI: $[-2.40, +2.60]$ percentage points). This result provides a crucial nuance to our initial hypothesis. Architectural diversity, as implemented in SEER-Full, remains valuable for mitigating model-specific biases, but it is not the primary driver of the observed performance gain. Instead, the evidence indicates that the main advantage of our framework stems from performing multiple extraction runs and aggregating their outputs with an LLM.

4.4.3. Ensemble size: 3-agent vs. 2-agent

Block C evaluates the marginal contribution of the third extractor by testing all three possible 2-agent pairs drawn from the SEER-Full ensemble. In each configuration, the excluded model serves as aggregator, ensuring a clean separation of roles across all three variants.

All three 2-agent configurations fall below the SEER-Full baseline, with drops ranging from -2.16% (C1: Kimi-K2+DS-V3.1) to -4.28% (C2: GLM-4.5+Kimi-K2). C1, which retains the two individually strongest extractors and omits GLM-4.5, produces the smallest degradation, yet still falls measurably below the three-agent baseline, indicating that GLM-4.5 contributes extractable signal even as the weakest individual model. C2, which omits DS-V3.1, produces the largest drop, suggesting that DS-V3.1 covers extraction patterns that are least overlapping with those of the other two models.

This pattern indicates that each model in the SEER-Full ensemble makes a distinct contribution to the final output, and that the three-agent configuration achieves a coverage that no 2-agent subset replicates.

4.4.4. Robustness to weaker agents

Block D evaluates the robustness of the ensemble when one top-tier extractor is replaced by a weaker model. In D1, GLM-4.5 is substituted with GPT-20b, whose single-agent F1 of 69.04% (Table 7) is approximately 4 points below that of GLM-4.5 (72.99%). Kimi-K2 and DS-V3.1 are retained as the two remaining extractors, and GLM-4.5 continues to serve as aggregator.

The resulting configuration achieves 77.14% F1, a drop of only -1.08% relative to SEER-Full. This degradation is substantially smaller

Table 14

Cascade entity linking evaluation. The same linker (openai/gpt-oss-20b) used for the isolated benchmark is applied to agents extracted by 11 claim extraction models on the cascade gold standard (87 agents, 12 articles). Δ measures the performance difference compared to the isolated baseline on this same cascade subset (90.80%).

Configuration	Micro F1	Δ vs. Isolated
Direct linker on Cascade GS (87 agents, 12 articles)	90.80%	–
<i>Cascade (Extraction → Linker)</i>		
SEER (Multi-Agent Ensemble) (GLM-4.5 + Kimi-K2 + DS-V3.1)	90.80%	±0.00%
openai/gpt-oss-120b	90.80%	±0.00%
z-ai/glm-4.6	89.66%	–1.14%
moonshotai/kimi-k2-0905	89.66%	–1.14%
qwen/qwen3-235b-a22b-2507	88.51%	–2.29%
openai/gpt-4.1-nano	88.51%	–2.29%
z-ai/glm-4.5	87.36%	–3.44%
openai/gpt-oss-20b	86.21%	–4.59%
deepseek/deepseek-chat-v3.1	85.06%	–5.74%
google/gemini-2.5-flash-lite	85.06%	–5.74%
meta-llama/llama-4-maverick	77.01%	–13.79%

than that observed when removing an extractor entirely in Block C (–2.16% to –4.28%), suggesting that a weaker model still provides useful signal to the aggregator, even if its individual output quality is lower.

Overall, this experiment indicates that the ensemble can tolerate the inclusion of a mid-tier extractor without substantial performance loss, suggesting a degree of robustness in the aggregation mechanism that may be relevant in cost-constrained deployment scenarios.

4.5. Cascaded entity linking evaluation

While the previous sections evaluated the Claim Extraction and Entity Linking modules in isolation, real-world deployments use them as a cascaded pipeline. Errors in the upstream extraction module, such as missed agents or noisy names, can therefore propagate to the downstream linking module. To quantify this effect, we conducted an end-to-end evaluation.

We first constructed a dedicated Cascade Gold Standard by sampling 12 articles from CED-1K, with three articles drawn from each of the four topics. From their gold-standard claims, we collected distinct agent mentions at the article level, yielding 87 unique agent-article pairs: 49 with a valid Wikidata ID and 38 without one. To establish a baseline for this subset, we ran the isolated linker (openai/gpt-oss-20b) directly on the 87 gold-standard agents. This evaluation yielded a micro-F1 score of 90.80%. The difference from the 94.44% reported in Section 4.3 reflects the different sample used for this evaluation.

For the cascade evaluation, the 11 extraction models first extracted claims and agents from the 12 articles, after which the extracted agents were processed by the same linker. We evaluated the resulting cascade outputs against the Cascade Gold Standard using fuzzy matching.

Table 14 reports the results. The SEER Multi-Agent Ensemble and gpt-oss-120b both achieve a micro-F1 score of 90.80%, matching the isolated baseline and showing no observed performance decline on this subset. Across all models, the mean change relative to the baseline is –3.35 percentage points, with larger reductions observed for lower-performing extractors, such as meta-llama/llama-4-maverick (–13.79 percentage points).

These results suggest that error propagation from extraction to entity linking remains limited when high-performing extractors are used. This robustness may reflect two factors: (1) the ensemble extractor identifies the relevant agents, thereby reducing upstream false negatives; and (2) the entity linking module accommodates minor variations and noise in agent names through its NER-based refinement and LLM-based reasoning steps.

4.6. Computational overhead and scalability

Processing news articles with LLM-based pipelines introduces computational overhead that scales with both the number of articles and the number of agents extracted per article. Since API pricing for LLMs is highly volatile, evaluating efficiency solely in terms of monetary cost would limit the long-term reproducibility of the study. We therefore formalise the computational overhead of SEER primarily in terms of token consumption, an intrinsic and stable measure of the system's architectural complexity. Crucially, because LLM providers apply substantially different pricing to prompt processing and generation, we model input and output tokens separately. We then derive practical monetary cost estimates as a snapshot analysis to support deployment planning.

4.6.1. Token consumption model

Let T_{prompt} and T_{article} denote the token counts for the system instructions and the source article text, respectively.

For a standard single-model approach, the claim extraction is performed via a single LLM call. The token footprint per article is strictly divided into input context and generated output:

$$T_{\text{in}}^{\text{single}} = T_{\text{prompt}} + T_{\text{article}} \quad (1)$$

$$T_{\text{out}}^{\text{single}} = T_{\text{claims}} \quad (2)$$

In contrast, the SEER Multi-Agent Ensemble divides the extraction workload into two distinct phases: parallel extraction and aggregation. First, three independent models (M_A, M_B, M_C) process the article. The token consumption for this phase is:

$$T_{\text{in}}^{\text{extractors}} = 3 \cdot (T_{\text{prompt}} + T_{\text{article}}) \quad (3)$$

$$T_{\text{out}}^{\text{extractors}} = \sum_{i \in \{A, B, C\}} T_{\text{claims}}^i \quad (4)$$

Second, the aggregator model synthesises these results. By design, the aggregator does not receive the original article text. This architectural choice enforces a strict synthesis role, preventing the aggregator from performing a redundant fourth extraction, and significantly reduces the input token overhead. Its input consists solely of a specific aggregator prompt ($T_{\text{agg_prompt}}$) and the concatenated outputs of the three extractors. Its output ($T_{\text{claims}}^{\text{final}}$) is a single, consolidated list of claims, which is structurally comparable in size to a single model's output ($T_{\text{claims}}^{\text{final}} \approx T_{\text{claims}}$). The aggregator's token footprint is therefore:

$$T_{\text{in}}^{\text{aggregator}} = T_{\text{agg_prompt}} + T_{\text{out}}^{\text{extractors}} \quad (5)$$

$$T_{\text{out}}^{\text{aggregator}} = T_{\text{claims}}^{\text{final}} \quad (6)$$

Table 15

Empirical token consumption per article for claim extraction and entity linking. Values represent averages measured over a stratified sample of 12 articles from the CED-1K evaluation corpus.

Configuration	Avg input tokens	Avg output tokens
Extraction phase		
Single Model (Kimi-K2)	2,846	1,920
SEER Ensemble (Multi-Agent)	15,011	20,954
Enrichment phase		
Entity Linking (Avg per Article)	29,718	3,807
Total pipeline footprint		
Single-Agent Pipeline (Kimi-K2 + EL)	32,564	5,727
SEER Full Pipeline (Ensemble + EL)	44,729	24,761

The total token consumption for the SEER claim extraction stage is therefore the sum of these two phases:

$$T_{in}^{ensemble} = T_{in}^{extractors} + T_{in}^{aggregator} \quad (7)$$

$$T_{out}^{ensemble} = T_{out}^{extractors} + T_{out}^{aggregator} \quad (8)$$

Finally, the Entity Linking stage introduces a variable overhead depending on the number of agents discovered in the text. For each extracted agent, SEER executes a maximum of two LLM calls. The first call, candidate resolution (*res*), is conditional: it is executed only if the initial Wikidata lookup returns one or more candidates (encompassing both validation and disambiguation). The second call, taxonomy classification (*tax*), is executed for all extracted agents regardless of the lookup outcome.

Let $\bar{N}_{ag}$ be the average total number of agents extracted per article, and let $\bar{N}_{res}$ be the average number of agents per article that require the resolution step ($\bar{N}_{res} \leq \bar{N}_{ag}$). The entity linking token footprint per article is accurately modelled as:

$$T_{in}^{el} = (\bar{N}_{res} \cdot T_{in}^{res}) + (\bar{N}_{ag} \cdot T_{in}^{tax}) \quad (9)$$

$$T_{out}^{el} = (\bar{N}_{res} \cdot T_{out}^{res}) + (\bar{N}_{ag} \cdot T_{out}^{tax}) \quad (10)$$

4.6.2. Empirical estimates and monetary cost insight

Table 15 reports the empirical token consumption measured across our evaluation corpus (CED-1K). We compare the best-performing single model (moonshotai/kimi-k2-0905) against the full SEER pipeline.

Monetary cost implications. To translate the architectural token footprint into a practical financial estimate, we apply the API pricing snapshot from Q1 2026 (as detailed in Table 5).¹⁴ Let P_{in} and P_{out} be the cost per million input and output tokens for a given model. The cost per article is calculated as:

$$Cost = \left(\frac{T_{in}}{10^6} \cdot P_{in} \right) + \left(\frac{T_{out}}{10^6} \cdot P_{out} \right) \quad (11)$$

Applying this formula, processing a standard article using the best-performing single model (Kimi-K2) costs approximately \$0.0065. The Entity Linking phase, which leverages the highly cost-efficient gpt-oss-20b model, adds an average of only \$0.0014 per article (processing an average of $\bar{N}_{ag} \approx 5$ agents per document). Therefore, a baseline **Single-Agent Pipeline** requires an estimated budget of \$7.92 per 1000 articles.

In contrast, deploying the **SEER Full Pipeline** (Multi-Agent Ensemble + Entity Linking) raises the estimated budget to \$52.50 per 1000

articles. It is important to note that the substantial increase in output tokens for the Ensemble (as seen in Table 15) does not reflect a bloated final extraction, but is primarily driven by *reasoning tokens*. Models within the ensemble, particularly z-ai/glm-4.5, heavily utilise internal Chain-of-Thought generation to resolve semantic ambiguities and reach consensus, with reasoning tokens accounting for over 60% of the ensemble's total output footprint.

This comparison reveals a clear architectural trade-off. The multi-agent framework costs approximately 6.6 times as much as the best single-model baseline (\$52.50 versus \$7.92 per 1000 articles), making the difference increasingly consequential at scale. Accordingly, our cost-benefit analysis indicates that a single high-performing model or a lower-cost ensemble of lightweight models offers a competitive and cost-effective solution for exploratory analyses and budget-constrained projects. By contrast, the SEER Full Pipeline is a specialised configuration for settings that require high extraction fidelity. Its higher cost is most clearly justified in applications with stringent reliability requirements, such as institutional fact-checking, rigorous academic media analysis, and legal auditing, where maximising extraction reliability and minimising hallucinations take precedence over budget constraints. Further reductions in LLM inference costs would lower the absolute financial premium associated with multi-agent architectures, potentially making this robust approach viable for a broader range of deployments.

4.6.3. Inference latency and scalability

Beyond token consumption, a critical factor for large-scale deployment is runtime overhead. While a multi-agent framework inherently requires more computation than a single-model approach, SEER mitigates inference latency through architectural design. The three extractor models (M_A, M_B, M_C) operate strictly in parallel. Consequently, the bottleneck for the extraction phase is not the sum of their execution times, but rather the maximum latency among the three extractors, plus the sequential execution of the single aggregator model. Similarly, during the entity linking phase, multiple extracted agents within the same article can be processed concurrently. This high degree of parallelisability ensures that while token volume increases, the wall-clock runtime scales efficiently, allowing the pipeline to process large document streams when deployed on infrastructure capable of concurrent API requests.

4.7. Downstream application: Large-scale viewpoint and sentiment analysis

To empirically demonstrate the practical analytical value of the Knowledge Graph generated by SEER, we designed a use case focused on media representation and political discourse analysis. Our aim is to show that the KG extracted by SEER supports complex aggregated queries that could not be executed directly on raw text.

We deployed SEER on a large-scale, unannotated corpus comprising 4K news articles focused on UK immigration and covering the period from 2023-02-20 to 2024-12-31. The system successfully extracted 39,188 individual claims. The Entity Linking module identified 13,354 agent mentions, resolving them into 6067 unique agents, of which 1581 were grounded to canonical Wikidata identifiers. This canonicalisation is a foundational step, as it allows thousands of disparate mentions to be aggregated into coherent actor profiles.

As a demonstration, we queried the resulting KG via SPARQL queries to isolate all claims attributed to six prominent UK political figures, deliberately selected to represent distinct ideological groupings: Yvette Cooper, Keir Starmer, and Roza Salih (representing the centre-left), and Rishi Sunak, Tom Pursglove, and James Cleverly (representing the centre-right). We further applied the LLM-based classification step from previous work [7,68] to map each extracted claim to a set of immigration viewpoints and assess its underlying sentiment. In this context, viewpoints are considered aggregations of

¹⁴ It is important to note that the monetary costs reported below reflect actual API billing during our experimental runs. These real-world figures incorporate dynamic API features, such as prompt caching and differential pricing for reasoning tokens, which are not captured by a simple multiplication of the average token counts (Table 15) by the base prices (Table 5).

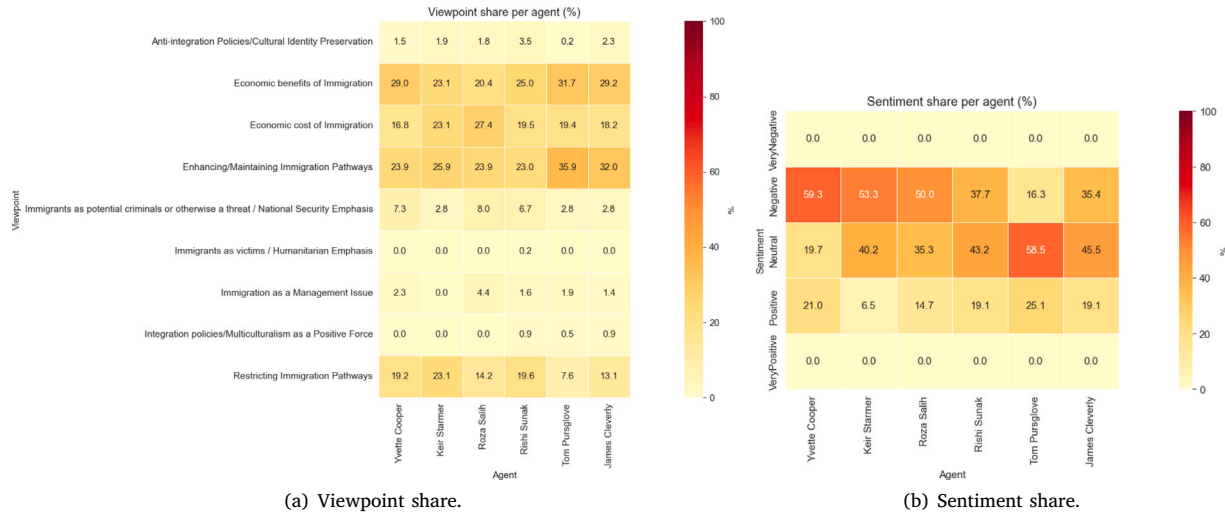


Fig. 6. Downstream analysis enabled by the SEER Knowledge Graph on a corpus of 4K articles about Immigration. (a) Distribution of viewpoints across claims made by six key political figures, grouped by ideological alignment (Center-Left vs. Center-Right). (b) Sentiment distribution of the extracted claims. The aggregation is made possible by the upstream entity linking and structural modelling provided by SEER.

claims that share the same perspective [10], making it possible to capture the different narratives circulating within a media sphere.

The results of this aggregation are visualised in Fig. 6. Fig. 6(a) illustrates the viewpoint share per agent, revealing how the KG captures distinct political positionings. For instance, while all actors frequently discuss the “Economic benefits of Immigration” and “Enhancing/Maintaining Immigration Pathways”, ideological nuances emerge in other categories. The Center-Left figures (Cooper and Starmer) exhibit a higher focus on “Restricting Immigration Pathways” (19.2% and 23.1% respectively) compared to the Center-Right ministers (Pursglove at 7.6% and Cleverly at 13.1%) in this specific corpus slice, highlighting the complex dynamics of political messaging.

Fig. 6(b), displaying sentiment share, provides further pragmatic insights into political communication. The Center-Left agents (Cooper, Starmer, Salih) exhibit a predominantly negative sentiment in their claims (ranging from 50.0% to 59.3%), consistent with a rhetorical posture focused on critiquing existing policies. Conversely, the Center-Right figures (Sunak, Pursglove, Cleverly) display a higher share of neutral sentiment (up to 58.5% for Pursglove), reflecting a more administrative or policy-announcement rhetorical style typical of executive roles.

This downstream application demonstrates the practical utility of SEER. By transforming unstructured journalistic text into a formally structured Knowledge Graph, the framework bridges the gap between raw media streams and high-level computational social science, enabling researchers and other stakeholders to systematically quantify viewpoint diversity, ideological divides, and media framing at scale.

5. Conclusions

In this paper, we propose SEER, a comprehensive framework for identifying claims in news articles, attributing them to the relevant persons or organisations, and constructing a large-scale KG that systematically represents the claims contained in a collection of articles. We also presented a prototype implementation of the proposed architecture and evaluated it on a corpus of articles drawn from four domains: climate change, UK immigration, the war in Ukraine, and US politics.

The central hypothesis of our methodology was that a multi-agent ensemble architecture can achieve superior performance compared to single LLM-based agents, while also reducing bias and factual errors. The evaluation demonstrates that SEER indeed achieves excellent performance in both agent identification (85.74% Macro F1) and claim

extraction (78.22% Macro F1), clearly outperforming ten alternative single-model LLM-baselines on the claim extraction task. Moreover, SEER proves to be highly robust across all four domains and across different types of utterance, including direct, indirect, and hybrid forms, whereas individual models are more susceptible to contextual variation. The entity linking mechanism implemented in SEER also outperforms a range of alternative models, as well as a baseline based on direct lookup in Wikidata.

The results of our research have several important implications for both the research community and industry. From a research perspective, this work provides empirical evidence that multi-agent architectures can significantly improve the performance and reliability of complex information extraction pipelines. It opens new directions for the study of agent specialisation, collaboration strategies, and conflict resolution mechanisms, as well as for scalable and bias-aware KG construction from unstructured text. The integration of claim extraction, attribution, and entity linking within a unified framework also enables further investigation into transparency, factual consistency, and trustworthiness in AI-driven knowledge acquisition systems. From an industry perspective, potential applications include media intelligence, fact-checking, policy analysis, reputation management, and the detection of misinformation, and lack of fairness and balance in media coverage. The demonstrated robustness across domains and utterance types makes the system suitable for real-world deployment in environments characterised by linguistic variability and contextual ambiguity. Furthermore, interoperability with existing resources such as Wikidata facilitates integration with enterprise data infrastructures. The multi-agent solution described here can therefore be adopted and integrated into existing systems for decision support in journalism, governance, and corporate strategy.

While the evaluation demonstrates strong performance across four domains, expanding the gold-standard corpus with additional articles, domains, and annotators will be an important step towards further validating the generalisability and robustness of the proposed approach. This study opens several promising avenues for future work. First, we intend to integrate the architecture for claim extraction with the methodology for viewpoint identification that we previously released [68], with the aim of producing a self-contained framework capable of delivering a comprehensive analysis of the viewpoints of multiple agents, organisations, and political and ideological parties at scale. Second, we plan to apply this technology to the analysis of a wide range of topics in order to map the current information

landscape in collaboration with colleagues in media studies. Third, we intend to further refine both the agentic architecture and the LLMs used for claim extraction and viewpoint identification, with the goal of addressing the current limitations discussed earlier. Finally, we plan to develop a fully open application that can be used by media studies experts and other users to produce in-depth analytics across multiple domains, countries, and news outlets. The ultimate goal is to facilitate the analyses of media representations of democratic discourse, thereby fostering a more transparent, inclusive, and resilient public debate.

CRedit authorship contribution statement

Massimiliano Fadda: Writing – original draft, Validation, Software, Methodology, Data curation. **Enrico Motta:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Francesco Osborne:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Diego Reforgiato Recupero:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Angelo Salatino:** Writing – review & editing, Software, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Aylien.com, a Quantexa solution, for providing us access to their news data. Moreover, this article is based on work from COST Action CA24121 KGELL and CA23147 GOBLIN, supported by COST (European Cooperation in Science and Technology).

Data availability

<https://github.com/massi-fa/SEER>.

References

- [1] S. Zaklama, Exploring the foundations of media framing theory, *Eur. Mod. Stud.* 9 (1) (2025) 75–89, [http://dx.doi.org/10.59573/emsj.9\(1\).2025.7](http://dx.doi.org/10.59573/emsj.9(1).2025.7).
- [2] M. Djerf-Pierre, A. Shehata, B. Johansson, Media salience shifts and the public's perceptions about reality: How fluctuations in news media attention influence the strength of citizens' sociotropic beliefs, *Mass Commun. Soc.* 28 (3) (2025) 459–484, <http://dx.doi.org/10.1080/15205436.2023.2299209>, [arXiv:https://doi.org/10.1080/15205436.2023.2299209](https://doi.org/10.1080/15205436.2023.2299209).
- [3] U. Reisch, The responsibility of social media in times of societal and political manipulation, *European J. Oper. Res.* 291 (3) (2021) 906–917, <http://dx.doi.org/10.1016/j.ejor.2020.09.020>, URL: <https://www.sciencedirect.com/science/article/pii/S0377221720308249>.
- [4] A. Bermes, Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumers' resilience during COVID-19, *J. Retail. Consum. Serv.* 61 (2021) 102555.
- [5] J.W. Boumans, D. Trilling, Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars, *Rethink. Res. Methods Age Digit. Journal.* (2018) 8–23.
- [6] D. Nguyen, How news media frame data risks in their coverage of big data and AI, *Internet Policy Rev.* 12 (2) (2023) 1–30.
- [7] E. Motta, F. Osborne, M.M.L. Pulici, A. Salatino, I. Naja, Capturing the viewpoint dynamics in the news domain, in: M. Alam, M. Rospocher, M. van Erp, L. Hollink, G.A. Gesese (Eds.), *Knowledge Engineering and Knowledge Management*, Springer Nature Switzerland, Cham, 2025, pp. 18–34.
- [8] A. Masini, P. Van Aelst, T. Zerback, C. Reinemann, P. Mancini, M. Mazzoni, M. Damiani, S. Coen, Measuring and explaining the diversity of voices and viewpoints in the news: A comparative study on the determinants of content diversity of immigration news, *Journal. Stud.* 19 (15) (2018) 2324–2343.
- [9] R. Schmälzle, R. Huskey, Integrating media content analysis, reception analysis, and media effects studies, *Front. Neurosci.* 17 (2023) 1155750.
- [10] E. Motta, E. Daga, A. Gangemi, M.L. Gjelsvik, F. Osborne, A. Salatino, The epistemology of fine-grained news classification, *Semant. Web* 16 (3) (2025) 22104968251344461, <http://dx.doi.org/10.1177/22104968251344461>, [arXiv:https://doi.org/10.1177/22104968251344461](https://doi.org/10.1177/22104968251344461).
- [11] F. Plötzky, W.-T. Balke, It's the same old story! Enriching event-centric knowledge graphs by narrative aspects, in: 14th ACM Web Science Conference 2022, in: *WebSci 201922*, ACM, New York, NY, USA, 2022, pp. 34–43, <http://dx.doi.org/10.1145/3501247.3531565>.
- [12] S. Pareti, T. O'Keefe, I. Konstantas, J.R. Curran, I. Koprinska, Automatically detecting and attributing indirect quotations, in: D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 989–999, URL: <https://aclanthology.org/D13-1101/>.
- [13] T. O'Keefe, S. Pareti, J.R. Curran, I. Koprinska, M. Honnibal, A sequence labelling approach to quote attribution, in: J. Tsujii, J. Henderson, M. Pa.sca (Eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 790–799, URL: <https://aclanthology.org/D12-1072/>.
- [14] D. Elson, N. Dames, K. McKeown, Extracting social networks from literary fiction, in: J. Hajič, S. Carberry, S. Clark, J. Nivre (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 138–147, URL: <https://aclanthology.org/P10-1015/>.
- [15] T. Ni, Q. Wang, G. Ferraro, Explore BiLSTM-CRF-based models for open relation extraction, 2021, *CoRR* abs/2104.12333. [arXiv:2104.12333](https://arxiv.org/abs/2104.12333). URL: <https://arxiv.org/abs/2104.12333>.
- [16] Y. Chen, Z.-H. Ling, Q.-F. Liu, A neural-network-based approach to identifying speakers in novels, in: *Interspeech 2021*, 2021, pp. 4114–4118, <http://dx.doi.org/10.21437/Interspeech.2021-609>.
- [17] Y. Zhang, Y. Liu, DirectQuote: A dataset for direct quotation extraction and attribution in news articles, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2022, pp. 6959–6966, URL: <https://aclanthology.org/2022.lrec-1.752/>.
- [18] D. Metropolitansky, J. Larson, Towards effective extraction and evaluation of factual claims, 2025, [arXiv:2502.10855](https://arxiv.org/abs/2502.10855). URL: <https://arxiv.org/abs/2502.10855>.
- [19] L. Zheng, C. Li, Z. Liu, F. Huang, H. Jia, Z. Ye, X. Zhang, Fact in fragments: Deconstructing complex claims via LLM-based atomic fact extraction and verification, 2025, [arXiv:2506.07446](https://arxiv.org/abs/2506.07446). URL: <https://arxiv.org/abs/2506.07446>.
- [20] P. Manakul, A. Liusie, M.J.F. Gales, SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models, 2023, [arXiv:2303.08896](https://arxiv.org/abs/2303.08896). URL: <https://arxiv.org/abs/2303.08896>.
- [21] Q. Liu, X. Chen, Y. Ding, B. Song, W. Wang, S. Wu, L. Wang, Attention-guided self-reflection for zero-shot hallucination detection in large language models, 2025, [arXiv:2501.09997](https://arxiv.org/abs/2501.09997). URL: <https://arxiv.org/abs/2501.09997>.
- [22] N.D. Cao, G. Izcard, S. Riedel, F. Petroni, Autoregressive entity retrieval, 2020, *CoRR* abs/2010.00904. [arXiv:2010.00904](https://arxiv.org/abs/2010.00904). URL: <https://arxiv.org/abs/2010.00904>.
- [23] T. Ayoola, S. Tyagi, J. Fisher, C. Christodoulopoulos, A. Pierleoni, ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking, 2022, [arXiv:2207.04108](https://arxiv.org/abs/2207.04108). URL: <https://arxiv.org/abs/2207.04108>.
- [24] Y. Eshel, N. Cohen, K. Radinsky, S. Markovitch, I. Yamada, O. Levy, Named entity disambiguation for noisy text, in: R. Levy, L. Specia (Eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning*, CoNLL 2017, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 58–68, <http://dx.doi.org/10.18653/v1/K17-1008>, URL: <https://aclanthology.org/K17-1008/>.
- [25] D. Mueller, G. Durrett, Effective use of context in noisy entity linking, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1024–1029, <http://dx.doi.org/10.18653/v1/D18-1126>, URL: <https://aclanthology.org/D18-1126/>.
- [26] Y. Ding, Q. Zeng, T. Weninger, ChatEL: Entity linking with chatbots, 2024, [arXiv:2402.14858](https://arxiv.org/abs/2402.14858). URL: <https://arxiv.org/abs/2402.14858>.
- [27] J. Hou, C. Zhang, R. Meng, Harnessing deep LLM participation for robust entity linking, 2025, [arXiv:2511.14181](https://arxiv.org/abs/2511.14181). URL: <https://arxiv.org/abs/2511.14181>.
- [28] Y. Li, A. Galimov, M.D. Ganapaneni, P. Thejaswi, D. Meng, P. Kumar, S. Potdar, Leveraging the power of large language models in entity linking via adaptive routing and targeted reasoning, 2025, [arXiv:2510.20098](https://arxiv.org/abs/2510.20098). URL: <https://arxiv.org/abs/2510.20098>.
- [29] J. Huang, X. Chen, S. Mishra, H.S. Zheng, A.W. Yu, X. Song, D. Zhou, Large language models cannot self-correct reasoning yet, 2024, [arXiv:2310.01798](https://arxiv.org/abs/2310.01798). URL: <https://arxiv.org/abs/2310.01798>.
- [30] Q. Zhang, D. Wang, H. Qian, Y. Li, T. Zhang, M. Huang, K. Xu, H. Li, Y. Liu, H. Qiu, Understanding the dark side of LLMs' intrinsic self-correction, 2025, [arXiv:2412.14959](https://arxiv.org/abs/2412.14959). URL: <https://arxiv.org/abs/2412.14959>.

- [31] Y. Du, S. Li, A. Torralba, J.B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, 2023, [arXiv:2305.14325](https://arxiv.org/abs/2305.14325). URL: <https://arxiv.org/abs/2305.14325>.
- [32] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, Z. Tu, Encouraging divergent thinking in large language models through multi-agent debate, 2024, [arXiv:2305.19118](https://arxiv.org/abs/2305.19118). URL: <https://arxiv.org/abs/2305.19118>.
- [33] L.B. Kaesberg, J. Becker, J.P. Wahle, T. Ruas, B. Gipp, Voting or consensus? Decision-making in multi-agent debate, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 11640–11671, <http://dx.doi.org/10.18653/v1/2025.findings-acl.606>.
- [34] Y. Liu, J. Cao, Z. Li, R. He, T. Tan, Breaking mental set to improve reasoning through diverse multi-agent debate, in: The Thirteenth International Conference on Learning Representations, 2025, URL: <https://openreview.net/forum?id=t6QHYUOQL7>.
- [35] Y. Ning, X. Lin, F. Fang, Y. Cao, MAD-fact: A multi-agent debate framework for long-form factuality evaluation in LLMs, 2025, [arXiv:2510.22967](https://arxiv.org/abs/2510.22967). URL: <https://arxiv.org/abs/2510.22967>.
- [36] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, W. Han, ChatIE: Zero-shot information extraction via chatting with ChatGPT, 2024, [arXiv:2302.10205](https://arxiv.org/abs/2302.10205). URL: <https://arxiv.org/abs/2302.10205>.
- [37] M. Lu, Y. Xie, Z. Bi, S. Cao, X. Wang, CROSSAGENTIE: Cross-type and cross-task multi-agent LLM collaboration for zero-shot information extraction, in: W. Che, J. Nabende, E. Shutova, M.T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 13953–13977, <http://dx.doi.org/10.18653/v1/2025.findings-acl.718>, URL: <https://aclanthology.org/2025.findings-acl.718/>.
- [38] Z. Wang, Z. Zhao, Y. Lyu, Z. Chen, M. de Rijke, Z. Ren, A cooperative multi-agent framework for zero-shot named entity recognition, 2025, [arXiv:2502.18702](https://arxiv.org/abs/2502.18702). URL: <https://arxiv.org/abs/2502.18702>.
- [39] W.R. van Hage, V. Malaisé, R. Segers, L. Hollink, G. Schreiber, Design and use of the simple event model (SEM), *J. Web Semant.* 9 (2) (2011) 128–136, <http://dx.doi.org/10.1016/j.websem.2011.03.003>, URL: <https://www.sciencedirect.com/science/article/pii/S1570826811000199>. Provenance in the Semantic Web.
- [40] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, *J. Web Semant.* 37–38 (2016) 132–151, <http://dx.doi.org/10.1016/j.websem.2015.12.004>, URL: <https://www.sciencedirect.com/science/article/pii/S1570826815001456>.
- [41] S. Gottschalk, E. Demidova, EventKG: A multilingual event-centric temporal knowledge graph, 2018, [arXiv:1804.04526](https://arxiv.org/abs/1804.04526). URL: <https://arxiv.org/abs/1804.04526>.
- [42] S. Angioni, S. Consoli, D. Dessì, F. Osborne, D.R. Recupero, A. Salatino, Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis, *IEEE Access* 12 (2024) 77269–77283.
- [43] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, ClaimsKG: A knowledge graph of fact-checked claims, in: The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 309–324, http://dx.doi.org/10.1007/978-3-030-30796-7_20.
- [44] F. Arslan, J. Caraballo, D. Jimenez, C. Li, Modeling factual claims with semantic frames, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2511–2520, URL: <https://aclanthology.org/2020.lrec-1.306/>.
- [45] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, 2022, [arXiv:2108.11896](https://arxiv.org/abs/2108.11896). URL: <https://arxiv.org/abs/2108.11896>.
- [46] E. Motta, E. Daga, A.L. Opdahl, B. Tessem, Analysis and design of computational news angles, *IEEE Access* 8 (2020) 120613–120626, <http://dx.doi.org/10.1109/ACCESS.2020.3005513>.
- [47] Y. Luan, L. He, M. Ostendorf, H. Hajishirzi, Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3219–3232, <http://dx.doi.org/10.18653/v1/D18-1360>, URL: <https://aclanthology.org/D18-1360/>.
- [48] D. Dessì, F. Osborne, D. Buscaldi, D. Reforgiato Recupero, E. Motta, Cs-kg 2.0: A large-scale knowledge graph of computer science, *Sci. Data* 12 (1) (2025) 964.
- [49] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, L. Secchi, Data-driven methodology for knowledge graph generation within the tourism domain, *IEEE Access* 11 (2023) 67567–67599.
- [50] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities, *World Wide Web* 27 (5) (2024) 58, <http://dx.doi.org/10.1007/s11280-024-01297-w>.
- [51] T. Aggarwal, A. Salatino, F. Osborne, E. Motta, Leveraging large language models for generating research topic ontologies: A multi-disciplinary study, *ACM Trans. Intell. Syst. Technol.* (2026).
- [52] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Knowledge graph validation by integrating LLMs and human-in-the-loop, *Inf. Process. Manage.* 62 (5) (2025) 104145, <http://dx.doi.org/10.1016/j.ipm.2025.104145>.
- [53] Y. Shao, X. Shi, Y. Zhang, N. Shiwakoti, Y. Xu, Z. Ye, Injury severity prediction and exploration of behavior-cause relationships in automotive crashes using natural language processing and extreme gradient boosting, *Eng. Appl. Artif. Intell.* 133 (2024) 108542, <http://dx.doi.org/10.1016/j.engappai.2024.108542>, URL: <https://www.sciencedirect.com/science/article/pii/S0952197624007000>.
- [54] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Trans. Knowl. Data Eng.* 36 (7) (2024) 3580–3599, <http://dx.doi.org/10.1109/tkde.2024.3352100>.
- [55] H. Bian, LLM-empowered knowledge graph construction: A survey, 2025, [arXiv:2510.20345](https://arxiv.org/abs/2510.20345). URL: <https://arxiv.org/abs/2510.20345>.
- [56] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, *Adv. Neural Inf. Process. Syst.* 35 (2022) 24824–24837.
- [57] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Commun. ACM* 57 (10) (2014) 78–85.
- [58] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Comput. Surv.* 55 (12) (2023) <http://dx.doi.org/10.1145/3571730>.
- [59] Qwen Team, Qwen3 technical report, 2025, [arXiv:2505.09388](https://arxiv.org/abs/2505.09388). URL: <https://arxiv.org/abs/2505.09388>.
- [60] OpenAI, gpt-oss-120b & gpt-oss-20b model card, 2025, [arXiv:2508.10925](https://arxiv.org/abs/2508.10925). URL: <https://arxiv.org/abs/2508.10925>.
- [61] Google DeepMind, Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025, [arXiv:2507.06261](https://arxiv.org/abs/2507.06261). URL: <https://arxiv.org/abs/2507.06261>.
- [62] OpenAI, Introducing GPT-4.1 in the API. URL: <https://openai.com/index/gpt-4-1/>. Weblog post.
- [63] Meta AI, The Llama 4 herd: The beginning of a new era of natively multimodal AI. URL: <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Weblog post.
- [64] DeepSeek-AI, DeepSeek-V3 technical report, 2024, [arXiv:2412.19437](https://arxiv.org/abs/2412.19437). URL: <https://arxiv.org/abs/2412.19437>.
- [65] GLM Team, A. Zeng, et al., GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025, [arXiv:2508.06471](https://arxiv.org/abs/2508.06471). URL: <https://arxiv.org/abs/2508.06471>.
- [66] Z.ai, GLM-4.6: Advanced agentic, reasoning and coding capabilities rivaling claude sonnet 4. URL: <https://z.ai/blog/glm-4.6>. Weblog post.
- [67] Kimi Team, Y. Bai, et al., Kimi K2: Open agentic intelligence, 2025, [arXiv:2507.20534](https://arxiv.org/abs/2507.20534). URL: <https://arxiv.org/abs/2507.20534>.
- [68] M. Fadda, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, Integrating large language models and knowledge graphs to capture political viewpoints in news media, *EPJ Data Sci.* 15 (1) (2026) 68, <http://dx.doi.org/10.1140/epjds/s13688-026-00678-7>.