Social Good or Scientific Curiosity?
Uncovering the Research Framing Behind NLP Artefacts

Eric Chamoun[1], Nedjma Ousidhoum[2]*, Michael Schlichtkrull[3],*, Andreas Vlachos[1]
[1]Department of Computer Science and Technology, University of Cambridge
[2]Cardiff University
[3]Queen Mary University of London
{ec806,av308}@cam.ac.uk, ousidhoumN@cardiff.ac.uk, m.schlichtkrull@qmul.ac.uk

Abstract

Clarifying the research framing of NLP artefacts (e.g., models, datasets, etc.) is crucial to aligning research with practical applications when researchers claim that their findings have real-world impact. Recent studies manually analyzed NLP research across domains, showing that few papers explicitly identify key stakeholders, intended uses, or appropriate contexts. In this work, we propose to automate this analysis, developing a three-component system that infers research framings by first extracting key elements (means, ends, stakeholders), then linking them through interpretable rules and contextual reasoning. We evaluate our approach on two domains: automated fact-checking using an existing dataset, and hate speech detection for which we annotate a new dataset[1] —achieving consistent improvements over strong LLM baselines. Finally, we apply our system to recent automated fact-checking papers and uncover three notable trends: a rise in underspecified research goals, increased emphasis on scientific exploration over application, and a shift toward supporting human fact-checkers rather than pursuing full automation.

1 Introduction

As NLP systems see wider real-world deployment, concerns over misuse and unintended consequences have intensified focus on responsible AI practices—highlighting the need for ethical integration, risk mitigation, and attention to societal impact (Noble, 2018; Bender et al., 2021; Birhane et al., 2022). A critical but often overlooked aspect of responsible NLP research is *research framing*—how a paper positions the purpose and value of its work by connecting *what* it does (means), *why* it does it (ends), and for *whom* it is relevant (stakeholders). In this context, framing highlights the methodological specification of an artefact's role within the research ecosystem. This contrasts with political or media framing, which emphasizes persuasion in discourse, and semantic framing, which concerns meaning in linguistic structure (Card et al., 2016).

When poorly articulated, research framings can obscure a system's scope, hinder accountability, and increase the risk of misuse. Figure 1 illustrates this point by comparing two NLP papers: one clearly intends to support professional fact-checkers with specified means, ends, and users; the other, on hate speech detection, lacks any mention of who would use the model's outputs. Such omissions risk misinterpretation—e.g., framing a label-only classification model as an automated decision tool overlooks the need for context and explanation, risking misuse in high-stakes settings.

This concern has motivated several analyses of how NLP research is framed. Schlichtkrull et al. (2023) first formalized the notion of research framings (which they termed "epistemic narratives") and introduced *epistemic elements*—textual signals that define a paper's framing by specifying components such as means, ends and stakeholders. Analyzing 100 automated fact-checking papers, they found frequent misalignments between goals and methods, underspecified or missing stakeholder definitions, and weak links between these elements. Separately, Liu et al. (2023) reviewed 333 summarization papers and found that fewer than 15% meaningfully addressed responsible AI concerns.

Such analyses help track field-wide trends to inform future research and reveal problems in research practices, including value misalignments and overlooked downstream risks (Birhane et al., 2022). They further offer guidance by encouraging alignment between research goals and stakeholder needs, prompting reflection on issues like dual use (Leins et al., 2020) and overclaiming (Grodzinsky et al., 2012). However, their reliance on manual annotation limits their scalability and timeliness.

---

*Equal contribution.
[1]Code and annotations available at our GitHub repository.

Assistive deployment (Fact-checking)	Vague deployment (Hate speech)
Excerpt: In an effort **to assist fact-checkers**, we tackle the **claim detection task**... **Misinformation** has recently become **more central in public discourse** ... NLP approaches to **alleviate** the activity of **fact-checking**. We introduce an approach ... **to perform the classification.**	**Excerpt:** Current filters are **insufficient to prevent the spread of hate speech.** Most **internet users**... report having been subjected to offensive name-calling... leverage our annotations to improve **hate speech detection**... **Treating hate speech classification** ... **label** around 13,000 potentially **derogatory tweets**.
Epistemic elements	**Epistemic elements**
• **Model means:** What (ML) methods are proposed? → classify/score veracity	• **Model means:** What (ML) methods are proposed? → classify/score posts/text, data collection
• **Application means:** How will the artefact be used? → identify claims	• **Application means:** How will the artefact be used? → identify toxic content
• **Data actors:** Who will act on the predictions from the artefact? → professional fact-checkers	• **Data actors:** Who will act on the predictions from the artefact? → not specified
• **Ends:** What is the intended societal goal? → fight misinformation	• **Ends:** What is the intended societal goal? → fight hate
Framing: Clear assistive framing, artefacts support journalists. → **Assisted external fact-checking**.	**Framing:** Clear goals and methods, but the user of the model's outputs is unspecified. → **Vague identification**.

Figure 1: Side-by-side examples of research framings. The fact-checking paper (left, Konstantinovskiy et al. (2021)) presents a clear assisted external fact-checking framing with specified users, goals, and methods. The hate speech paper (right, Ousidhoum et al. (2019)) has an underspecified framing due to unspecified model users.

In this work, we first propose a generalizable research framing schema to support cross-domain analysis in NLP (§2). Building on this, we introduce an automated framework for inferring research framings (§3), composed of three stages:

1. *Epistemic element extraction*, i.e., identifying parts of the introductory text that specify the components of a paper's research framing, such as its aims, means and stakeholders.

2. *Research framing ranking*, which uses semi-automatically inferred rules to connect extracted elements, linking means to ends to infer the artefact's intended purpose;

3. *Research framing classification*, which leverages an LLM to refine rankings with broader context, in-context examples, and reasoning.

We evaluate our approach (§5) in two domains: automated fact-checking (AFC) and hate speech detection (HS). For AFC, we use the dataset released by Schlichtkrull et al. (2023); for HS, we manually annotate 49 highly cited papers (§4) with epistemic elements, which we use to infer their research framings. Our approach outperforms a strong LLM baseline in both domains, offering more accurate research framing predictions. It also achieves high precision in identifying underspecified framings with misaligned or incomplete elements.

Finally, we apply our system (§6) to recent AFC papers to uncover emerging trends. We observe a rise in underspecified research framings, raising concerns about the clarity of current framing prac-

tices in AFC research. We also find an increase in papers categorized under *Scientific Curiosity*, largely driven by benchmark studies examining LLM limitations. Lastly, we find a growing emphasis on generating fact-checking justifications to assist human fact-checkers, while fewer studies propose fully automating the process.

Research framings for this paper We release two NLP artefacts: (1) datasets used for analysis and (2) a tool for generating research framing labels and justifications (*modeling means*). Both are designed for use by researchers (*data actors, model owners*), who are also the *data subjects*. The tool is intended to support large-scale analyses of NLP subfields and provide feedback on how artefacts are positioned for use or exploration (*application means*). The study aims to advance NLP knowledge and promote framing transparency (*ends*), aligning with the research framings of *Scientific Curiosity* and *Scientific Writing Assistance*.

2 Cross-domain research framing

A research framing captures how a paper connects key epistemic elements to convey the intended purpose of its contributions. We consider the following epistemic elements: (1) *data subjects*—individuals whose behaviors are analyzed (e.g., social media users), (2) *data actors*—those intended to use the model outputs (e.g., moderators), (3) *model owners*—entities controlling deployment (e.g., social media companies), (4) *modeling means*—the machine learning approach (e.g.,

General Framing	Description	AFC	HS
Automated deployment	System replaces a human task with minimal intervention.	Automated external fact-checking	Automated content moderation
Assistive deployment	System supports human decision-making.	Assisted internal/external fact-checking	Assisted content moderation
Knowledge access and curation	Organizes/synthesizes knowledge for future use.	Assisted knowledge curation	Assisted knowledge curation
Knowledge exploration	Explores models or data without specific application goals.	Scientific curiosity	Scientific curiosity
Governance	Supports legal, institutional, or compliance goals.	Law enforcement	Law enforcement
Vague deployment	Implies deployment but omits how or where the model is used.	Vague debunking	Vague moderation
Vague opposition	States a broad goal, but lacks a coherent link between that goal and the proposed ML method.	Vague opposition	Vague opposition
Vague identification (Detection tasks)	Identifies content without specifying how it is used to achieve stated ends and who acts on it.	Vague identification	Vague identification

Table 1: Cross-domain research framings, with corresponding instantiations in AFC, HS. Clear framings (blue) link technical contributions to explicit purposes/users; underspecified framings (pink) lack specificity about use/users/goals.

classification), (5) *application means*—how the model is used (e.g., automated removal), and (6) *ends*—the intended societal or research goal (e.g., fighting hate online). These elements (listed in Appendix A) may be explicitly connected, partially specified, or mismatched—e.g., presenting a classifier as a replacement for human experts without accounting for justification or user context.

To support automated analysis across domains, we propose a generalizable research framing schema. Table 1 presents our cross-domain framing types, along with their instantiations in AFC (Schlichtkrull et al., 2023) and HS (this work). While our evaluation focuses on these two domains, the schema is designed to be extensible; e.g., mental health applications like *automated monitoring* or *mandatory flagging* can align with this structure.

To assess framing quality, we distinguish between *clear* and underspecified cases. Clear framings specify the artefact's purpose, users, and goals; underspecified ones omit key components or introduce inconsistencies that obscure the system's intended use. We define underspecification relative to how clearly authors articulate and connect epistemic elements at the time of writing. In other words, it reflects presentation clarity in its contemporary context, not compliance with future or evolving community standards. This time-bounded definition makes the schema generalizable, since it provides a consistent basis for analyzing papers across different publication periods.

We also note that theoretical research without an immediate ends or intended user is not the same as underspecification. If a paper clearly states that the goal is theoretical exploration or to answer some research question, that corresponds to **Knowledge exploration**, a clear framing. Underspecification arises when this intent is unstated, e.g., authors claiming that their theoretical findings aim to *"fight hate"* without discussing how.

We highlight hate speech-specific framings (see Appendix A.1 for those specific to AFC):

LLM safety represents a clear framing focused on preventing harmful LLM outputs through detoxification, filtering, or alignment. **Vague data analysis** refers to papers combining knowledge exploration with social aims (e.g.,"fighting hate"), without a clear link between analysis and real-world impact.

3 Inference of Research Framings

We automatically infer the *research framings* in NLP papers in three stages (Figure 2):

3.1 Step 1: Epistemic Element Extraction

We treat element extraction as a multi-label classification task, since a paragraph may contain multiple relevant elements (e.g., several stakeholders or goals). We also allow the model to abstain when no evidence is present for any element category.

Formulation. Given a paragraph p_i from a selected paper and an epistemic element $E_j \in \mathcal{E}$ (e.g. *data actors*), we define the corresponding candidate label set $\mathcal{R}(E_j)$ (including *not specified*). We frame the extraction as a QA task, using a language model M to compute response likelihoods:
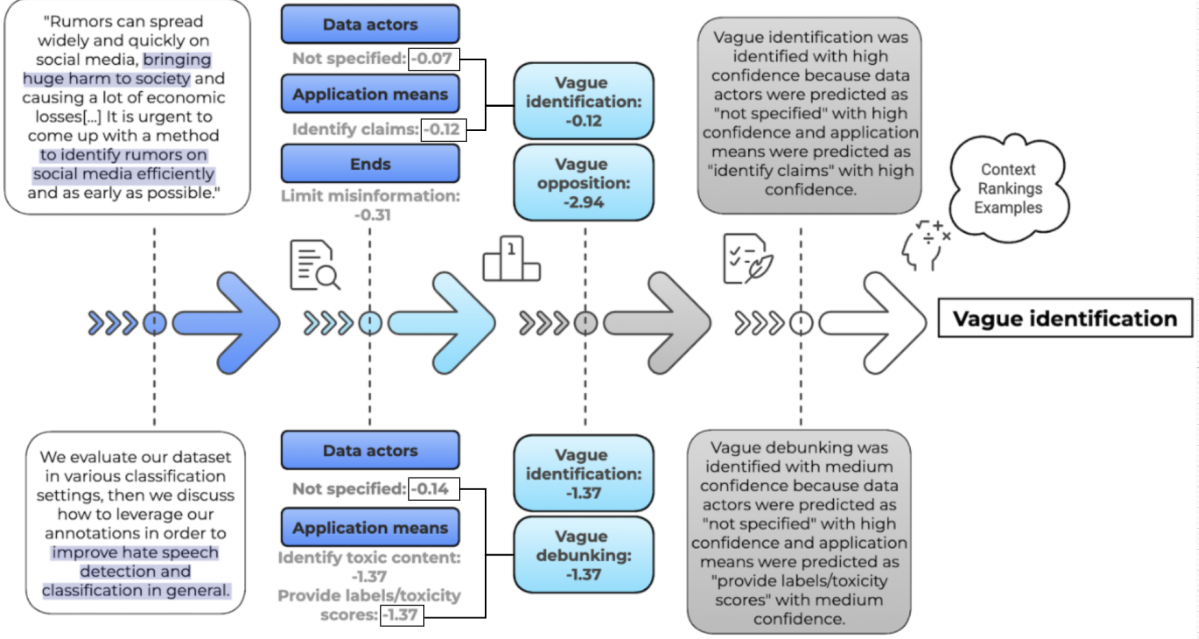
Figure 2: Our system infers a paper's research framings by: (1) estimating epistemic element likelihoods; (2) applying logical rules linking elements to framings; (3) generating explanations summarizing the likelihoods computed in steps 1 and 2 as reasoning for classification; (4) using an LLM to refine predictions using the explanations, context, and in-context examples. Numbers represent log-likelihoods of predicted elements or framings. Examples are from Bian et al. (2020) (top) and Ousidhoum et al. (2019) (bottom).

$$s_{p_i, r_{E_j,k}} = \log P_M(r_{E_j,k} \mid p_i, q_{E_j}) \quad (1)$$

where q_{E_j} is a predefined prompt targeting element E_j (e.g., *"Who are the individuals or entities that the authors identify as expected to act on the released artefacts?"*) and $r_{E_j,k} \in \mathcal{R}(E_j)$ is a candidate label (e.g. *journalists*).

Semantic Clustering via NLI. We adapt Kuhn et al. (2023) to cluster sampled outputs $\{x_1, \ldots, x_n\}$ into semantically equivalent classes $C = \{c_1, \ldots, c_m\}$. Two outputs x_a, x_b are grouped together if they entail each other bidirectionally (i.e., NLI returns entail in both directions). Class likelihoods are computed via:

$$s_{c_j} = \log \sum_{x \in c_j} \exp(s_{p_i,x}) \quad (2)$$

3.2 Step 2: Research Framing Ranking

In this step, we estimate the likelihood of each research framing $f_i \in \mathcal{F}$ by linking predicted epistemic elements through interpretable logical rules. Each research framing f_i is defined by a boolean rule $\mathcal{L}_i : \{r_{E_1,k_1}, r_{E_2,k_2}, \ldots, r_{E_m,k_m}\} \implies f_i$ that specifies how combinations of epistemic elements must co-occur to support that framing.

These rules are constructed from two sources: (1) a CART-based decision tree (Breiman et al., 1984) trained on annotated data, and (2) manual refinement based on framing definitions (Appendix C.1). An example is shown in Figure 3 in the appendix.

To compute the likelihood of a research framing, we evaluate each logical rule $\mathcal{L}_i$ over the predicted probabilities of its constituent epistemic elements. Conjunctions and disjunctions are computed as

$$\Pr(r_{E_a,k} \wedge r_{E_b,l}) = \min\{\Pr(r_{E_a,k}), \Pr(r_{E_b,l})\},$$

$$\Pr(r_{E_a,k} \vee r_{E_b,l}) = \max\{\Pr(r_{E_a,k}), \Pr(r_{E_b,l})\}.$$

. This soft logic enables smooth aggregation of partial evidence across elements.

At inference time, we aggregate paragraph-level predictions to compute paper-level scores for each epistemic element. Since research framings describe the broader intended purpose of an NLP artefact, these elements must be assessed across the full document rather than in isolated paragraphs.

Let a paper consist of paragraphs $\{p_1, p_2, \ldots, p_T\}$, and let $\Pr(r_{E_j,k} \mid p_t)$ denote the confidence score of label $r_{E_j,k}$ for element E_j in paragraph p_t. We compute the paper-level score as:

$$\Pr(r_{E_j,k}) = \max_{t=1}^{T} \Pr(r_{E_j,k} \mid p_t).$$

To model absence, we define the paper-level score for the *not specified* label of element E_j as:

$$\Pr(r_{E_j,\text{ns}}) = \min_{t=1}^{T} \Pr(r_{E_j,\text{ns}} \mid p_t).$$

This aggregation ensures that element presence is acknowledged if mentioned anywhere, while absence requires consistent omission across text.

3.3 Research Framing Classification

To refine framing predictions, we use an LLM M to incorporate document context and reason beyond deterministic rules. We first collect confidence scores from previous stages. Each framing f_i is associated with a confidence score $\Pr(f_i)$, and its supporting epistemic element labels S_i, as specified by $\mathcal{L}_i$, have scores $\Pr(r_{E_j,k})$. We discretize each score into a confidence tier for each paper:

$$\text{Tier}(\Pr(z)) = \begin{cases} \text{High} & \text{if } \Pr(z) \geq \tau_h \\ \text{Medium} & \text{if } \tau_m \leq \Pr(z) < \tau_h \\ \text{Low} & \text{if } \tau_l \leq \Pr(z) < \tau_m \end{cases}$$
$$\text{for all } z \in \{f_i\} \cup S_i$$

where τ_h, τ_m, τ_l are predefined confidence thresholds. We then generate structured natural language justifications summarizing the prediction rationale: *"[Framing] was identified with [confidence level] because [element 1] was predicted with [confidence level X], [element 2] with [confidence level Y], ..."*

M is fed explanations along with paper context, framing definitions, and examples (Appendix F).

4 Research Framings in Hate Speech

To assess cross-domain generalization, we apply our approach to hate speech detection—a task often framed as NLP for Social Good (Jin et al., 2021), where artefacts aim to assist moderators in combating online harm. We construct a new dataset of 49 highly cited papers on HS and related tasks from top NLP and CSS venues, annotated for epistemic elements and discourse-level research framings.

Annotation Procedure. We follow a two-phase protocol based on Krippendorff (2018):

1. Initial annotation used a predefined taxonomy of epistemic elements and research framings inspired by prior work (Schlichtkrull et al., 2023), with new categories added inductively as needed. Annotators met regularly to discuss and reconcile divergent cases.

2. A second round re-annotated all papers using the unified, finalized label set. No new labels were added in this phase, ensuring convergence on a minimal, interpretable set of research framing types.

Annotations were performed by two NLP researchers with domain expertise in hate speech and sociotechnical NLP. Inter-annotator agreement for the final label set reached a Krippendorff's α of 0.61 across the research framing labels (10 distinct research framings with multiple per paper).

Annotation Scope. For each paper, we extract quotes from the abstract and introduction that describe the system's purpose, methods and stakeholders. We annotate the presence of epistemic element types defined in §2. Each paper is then assigned one or more discourse-level *research framings*. Appendix E.3 defines the framing types. All examples were annotated based on explicitly stated text only, avoiding inferred intent. Detailed guidelines are provided in Appendix E.

Findings The distribution of research framings in our annotated HS dataset reveals that *Vague Identification* is the most prevalent (44.74%). These papers present systems for detecting toxic content but fail to specify who will act on the outputs or how. Overall, underspecified research framings make up 64.47% of the dataset, underscoring a broader trend of ambiguity in the intended use of hate speech detection models (see Appendix D for distribution details). This is reinforced by the prevalence of *Vague Data Analysis*, where papers combine knowledge exploration with social aims (e.g., "fighting hate") without articulating how the analysis supports real-world impact.

The second most common framing is *Scientific Curiosity*, often seen in papers that examine the linguistic characteristics of hate speech—such as nuance and sarcasm—or that introduce new datasets and benchmarks. Notably, only around 15% of papers clearly specify practical applications for mitigating hate speech, with a small percentage (2.63%) focusing specifically on limiting toxic content generated by LLMs.

5 Experimental setup

5.1 Implementation details

Modeling We use `gemini-1.5-flash-002` (Gemini Team et al., 2023) for all prompting tasks, including epistemic element extraction (Step

1) and research framing classification (Step 3). Gemini has demonstrated strong performance on open-ended tasks like writing assistance, balancing diversity and contextual alignment (Gooding et al., 2025). We provide all full prompts in Appendix F.

Epistemic Element Extraction For each paragraph, we generate $k = 10$ samples per epistemic element, following Kuhn et al. (2023), as increasing samples improves performance (Appendix B.1). We set the temperature to 1.0, balancing diversity and coherence—lower values produce overly deterministic responses, while higher ones risk incoherence. Clustering is performed using NLI with DeBERTa (He et al., 2020) to merge semantically equivalent responses. Entailment is determined by the model's predicted label.

Framing Ranking To map epistemic elements to research framings, we employ two decision trees: (1) a global tree that identifies key elements distinguishing research framings and (2) individual trees per research framing, capturing characteristic element combinations in a multi-label setting. Nodes are split using Gini impurity, with the global tree fully expanded for differentiation and individual trees capped at depth 4 for interpretability. Trees are trained using Gini impurity as the splitting criterion. The resulting paths are converted into boolean rules (see Section 3.2 and Appendix C.1).

Absence Thresholding Papers often present multiple research framings across sections—e.g., one section may frame a tool as identifying misleading claims without specifying intended actors (*Vague Identification*), while another specifies its use by *journalists*. Strictly assigning -∞ to *not specified* labels when *data actors* appear in some paragraphs would incorrectly exclude *Vague Identification*, even if applicable elsewhere. To balance specificity with flexibility, we set a -3 threshold for *not specified* labels—low enough to avoid inflating irrelevant elements (as model-assigned likelihoods exceed -3) but high enough to retain research framings dependent on their absence.

Framing Classification We define confidence tiers for both epistemic element and framing scores using the thresholds $\tau_h = -1$, $\tau_m = -2$, and $\tau_l = -3$, consistent with Gemini's typical scoring range of approximately $[-3, 0]$. These tiers are used to guide natural language justifications, which are passed to the LLM in conjunction with paper context, framing definitions, and in-context examples (see Figure 2). Using natural language labels for confidence levels improves the model's reasoning by making the signal more interpretable.

5.2 Evaluation metrics

We evaluate performance for **epistemic element extraction** and **research framing ranking** using Filtered Mean Reciprocal Rank (fMRR; Bordes et al. (2013)), which averages the inverse rank of correct predictions while excluding higher-ranked known positives. That is, when multiple correct answers exist, we compute the rank of a target answer after removing other correct answers ranked above it. Also, some labels are unscored ($-\infty$) if not predicted (e.g., *professional journalists* when only *social media users* are mentioned in the paper).

For scored labels, we compute standard fMRR. If a correct label is unscored, we estimate its rank as the number of scored labels plus half the unscored ones, assuming unscored items are randomly distributed. For example, with 2 scored and 10 unscored labels, the estimated rank is $2 + \frac{10}{2} = 7$, preventing artificially inflated results.

$$\text{fMRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i} \quad (3)$$

$$\text{rank}_i = \begin{cases} r_i, & \text{if label } i \text{ is scored} \\ s + \frac{u}{2}, & \text{if label } i \text{ is unscored} \end{cases} \quad (4)$$

where N is the number of evaluated instances, r_i the observed rank of the correct label, s the number of scored labels, u the number of unscored labels. For **research framing classification**, we report the Jaccard score, i.e., the intersection over union of predicted and gold labels: $\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$ (Jaccard, 1901). It is well-suited to our multi-label setting, where few labels are typically present among 10+ candidates. Unlike F-score—which can be inflated by correctly predicting absent labels—Jaccard evaluates only active labels. We also report F-score for specific framings for completeness.

5.3 Epistemic element extraction

We evaluate our semantic clustering-based approach for estimating epistemic element likelihoods in AFC and HS papers. Since our approach uses generation rather than classification, the model often produces semantically similar but syntactically varied outputs (e.g., "not mentioned," or "absent" when the prediction should be "not

| | AFC | | | | HS | |
| | Validation | | Test | | | |
	All	Oracle	All	Oracle	All	Oracle
Data Subjects	73.80	72.86	**86.19**	81.38	77.74	77.30
Data Actors	76.57	**77.87**	**77.44**	75.23	**76.13**	72.19
Model Owners	**96.64**	95.97	**97.77**	96.05	**96.60**	96.38
Model Means	64.11	**71.16**	56.21	65.79	**69.84**	68.99
App. Means	50.92	**55.14**	46.00	50.90	**66.18**	60.09
Ends	**76.58**	65.26	**78.96**	66.08	**74.04**	59.51
Total	**73.10**	73.04	**73.76**	72.57	**76.76**	72.41

Table 2: fMRR for epistemic elements on AFC and HS data. *All* uses all paragraphs from abstracts and introductions; *Oracle* uses paragraphs annotated by humans as containing research framing information.

Dataset	Approach	AFC	HS
Validation	System$_{all}$	55.51	-
	System$_{oracle}$	**65.04**	-
Test	System$_{all}$	**52.10**	52.17
	System$_{oracle}$	51.93	50.50
Entire Dataset	Human	84.84	83.63

Table 3: fMRR for research framing prediction across AFC and HS papers. System$_{all}$ aggregates epistemic element scores from all paragraphs, while System$_{oracle}$ uses only those labeled by humans as containing research framing information for ranking.

specified"). Without alignment, these would be treated as distinct labels. NLI-based clustering addresses this by grouping semantically equivalent outputs—thus supporting effective uncertainty quantification across multiple generations. As shown in Appendix B.1, performance improves with more samples, highlighting the benefits of our approach over a single-prompt baseline.

Table 2 shows that using all paragraphs slightly outperforms human-filtered ones, suggesting broader retrieval enhances epistemic element extraction, especially given the model's strong ability to filter irrelevant content (~76% fMRR for *not specified* labels). The model achieves near-perfect performance in identifying *model owners*, largely due to the high prevalence of *not specified* labels—fewer than 10% of paragraphs explicitly mention one. In contrast, performance on *application means* is lower, likely due to the greater complexity introduced by at least 10 distinct labels within this category, making classification more challenging. Despite these variations, the model performs consistently well across other epistemic elements. Performance is also consistent across domains, except for higher accuracy in identifying *application means* in hate speech detection. This may stem from AFC's greater label diversity (15 vs. 10), which includes both nuanced tasks (e.g., triaging claims) and concrete actions (e.g., automated removal), the latter shared across both domains.

5.4 Research framing ranking

Table 3 presents fMRR scores for research framing prediction across AFC and HS datasets. Despite the complexity of the task—10+ distinct research framings with multiple per paper—both approaches perform well. The upper-bound human performance (84.84% for AFC, 83.63% for HS), obtained by applying logical rules to manually annotated epistemic elements, indicates that logical mappings successfully capture relationships between epistemic elements and research framings.

5.5 Research framing classification

Table 4 presents research framing classification results. The baseline LLM, provided with in-context examples, research framing definitions, and paper context, achieves Jaccard scores of 21.89% (validation) and 28.45% (test) for AFC, and 16.79% for HS. This highlights the task's complexity, as it requires synthesizing multiple epistemic elements and reasoning over their connections. Augmenting the LLM with ranking-based explanations significantly improves performance, yielding gains of 6–7% in AFC and nearly 15% in HS . This suggests that structured guidance through deterministic research framing ranking explanations enhances inference, particularly in hate speech detection.

| | AFC | | HS |
	Validation	Test	
Baseline	21.89 ± 2.08	28.45 ± 1.40	16.79 ± 1.27
System	**30.67 ± 1.44**	**34.86 ± 1.05**	**31.37 ± 1.70**

Table 4: Jaccard score for research framing classification with 95% confidence intervals over 15 runs. *Baseline* uses an LLM with in-context examples, research framing definitions, and context, while *System* builds on this by incorporating rankings from our pipeline.

Table 5 reports per-research framing F-scores alongside ranking scores from Step 2. Performance gains strongly correlate with fMRR—higher ranking accuracy leads to greater classification. Only two research framings with sufficient samples show declines: *Assisted Media Consumption* (AFC) and *Vague Data Analysis* (HS). A closer examination reveals that deterministic rules struggle with context-dependent research framings like *Vague Data Anal-*

General Framing	Domain Framing	AFC				HS			
		C	fMRR$_{sys}$ (Step 2)	F1$_{base}$ (Step 3)	F1$_{sys}$ (Step 3)	C	fMRR$_{sys}$ (Step 2)	F1$_{base}$ (Step 3)	F1$_{sys}$ (Step 3)
Automated deployment	Automated external fact-checking	17	41.56	35.14 ± 2.70	**50.50** ± 2.28	–	–	–	–
	Automated content moderation	–	–	–	–	5	51.67	10.27 ± 7.01	**27.97** ± 9.41
Assistive deployment	Assisted external fact-checking	17	35.93	44.11 ± 2.12	**56.79** ± 2.34	–	–	–	–
	Assisted internal fact-checking	2	56.25	12.22 ± 12.94	26.48 ± 10.99	–	–	–	–
	Assisted content moderation	–	–	–	–	4	100.00	19.71 ± 1.31	**24.99** ± 2.84
Knowledge curation	Assisted knowledge curation	6	91.67	40.21 ± 4.73	**62.17** ± 3.46	–	–	–	–
Knowledge exploration	Scientific curiosity	5	52.41	18.92 ± 4.26	**20.24** ± 2.53	16	47.43	34.60 ± 3.62	**48.84** ± 2.96
	Adversarial research	1	50.00	46.00 ± 6.09	**72.67** ± 10.95	–	–	–	–
Governance	Law enforcement	1	100.00	100.00 ± 0.00	100.00 ± 0.00	–	–	–	–
Other	Assisted media consumption	4	36.58	**26.23** ± 4.50	18.92 ± 3.02	–	–	–	–
	Automated content moderation	1	14.29	**18.16** ± 6.93	0.00 ± 0.00	–	–	–	–
	LLM safety	–	–	–	–	2	100.00	55.56 ± 4.12	**63.11** ± 7.53
Vague deployment	Vague debunking	11	71.27	18.17 ± 5.89	**31.86** ± 1.63	–	–	–	–
	Vague moderation	–	–	–	–	7	41.90	3.64 ± 3.81	**14.80** ± 5.69
Vague identification	Vague identification	25	54.82	23.86 ± 4.64	**44.97** ± 2.01	34	51.30	2.92 ± 2.85	**35.79** ± 1.64
Vague opposition	Vague opposition	13	21.50	2.23 ± 2.34	**4.35** ± 3.23	0	–	–	–
Other	Vague data analysis	–	–	–	–	8	24.94	**61.53** ± 3.35	44.91 ± 4.03
Underspecified framing		49	–	34.70 ± 4.07	**58.91** ± 2.21	49	–	63.15 ± 2.50	**69.93** ± 2.70

Table 5: Research framing ranking and classification results in AFC and HS. **C**: Count. **fMRR$_{sys}$**: fMRR reporting system performance for ranking (step 2). **F1$_{base}$**: classification with baseline LLM. **F1$_{sys}$**: classification with baseline LLM incorporating rule-based rankings. 95% confidence intervals are from 15 runs.

ysis, which require reasoning about missing connections between means and ends. The LLM mitigates this by integrating context and reasoning to differentiate overlapping research framings. Nonetheless, most research framings benefit significantly from ranking-based explanations, with *Vague Identification* improving by over 20% in AFC and over 30% in HS. Clear research framings (top) also show important gains.

To evaluate the impact of confidence levels in research framing ranking justifications on the LLM's final classification, we analyze research framing retention rates. We find that 85.26% of research framings with high likelihoods (-1 to 0) are included in the final predictions, while 44.64% of medium-likelihood research framings are retained. In contrast, only 8.47% of the low-likelihood research framings are kept. This suggests that while the LLM considers confidence scores, it does not rely on them entirely and applies its own reasoning in the final classification. Finally, the system excels at detecting underspecified research framings, achieving F-scores of 58.91% (AFC) and 69.93% (HS) with precision rates of 76.16% and 86.24%.

6 Automated analysis of AFC literature

We apply our system to 102 automated fact-checking papers published between January 2023 and December 2024, to identify emerging trends (see Appendix D for distribution details).

Compared to Schlichtkrull et al. (2023), we observe a rise in *Scientific Curiosity* framings, driven by the growth of papers analyzing LLM capabilities in fact-checking. The model classifies works like Hu et al. (2024)—focused on evaluating LLM factual reasoning—as scientific curiosity, noting: *"The primary goal is to understand LLM limitations, a purely scientific endeavor."* Papers introducing datasets or benchmarks are similarly grouped under *Scientific Curiosity*, reflecting an emphasis on knowledge development over deployment.

We also observe a rise in *Vague Debunking* framings and a decline in *Vague Opposition*. The former describes papers that propose automated debunking without clarifying whether models support or replace human fact-checkers; the latter presents a broad goal without specifying how model outputs are used to achieve it. While our system reliably detects underspecification, it sometimes defaults to

Vague Debunking due to its broader scope—e.g., in Chamoun et al. (2023), it correctly identifies ambiguity in deployment but infers intended use based on model design, leading to misclassification.

Furthermore, we observe a shift from fully automated fact-checking (22.0% in Schlichtkrull et al. (2023) vs. 7.8% in our study) toward human-assisted systems (18.0% vs. 27.5%), reflecting growing awareness of LLM limitations. At the epistemic element level, *supplant human fact-checkers* drops sharply (24.0% $\rightarrow$ 3.9%), while *identifying misleading claims* for human review rises (36.0% $\rightarrow$ 55.9%). More papers aim to *generate justifications* (9.0% $\rightarrow$ 21.6%), supporting the trend of AI as assistive, not autonomous.

7 Conclusion

In this work, we first proposed a generalizable research framing schema to support consistent interpretation of research intent across NLP domains, then introduced an automated framework for extracting epistemic elements and inferring framings. To assess generalization, we annotated 49 HS papers and found our system outperforms a strong in-context LLM baseline in both AFC and HS. We applied the framework to 102 recent AFC papers to demonstrate its use for scalable literature analysis—revealing a shift toward human-assisted systems, a rise in underspecified framings, and greater emphasis on justification generation. Future work will explore automated framing and epistemic element discovery in new domains to further streamline NLP research analysis.

Acknowledgments

Limitations

While our study introduces an automated framework for meta-analyses in NLP research, we acknowledge the following:

- **Scope of defined research framings** – The predefined research framings, derived manually by domain experts from several papers, provide a strong foundation. However, they may not fully capture the evolving perspectives and emerging trends within the field.

- **Generalizability of conclusions** – Our findings reveal important trends, such as the prevalence of underspecified research framings. However, given that our analysis does not cover all NLP papers on automated fact-checking and hate speech detection, these insights should be viewed as well-informed estimates—not definitive claims. Nonetheless, they offer a strong starting point for understanding broader trends in research framing.

- **Tool accuracy** – While our model is effective at identifying underspecified research framings, achieving a precision of 76.16% in fact-checking and 86.24% in hate speech detection, we acknowledge that, as any other automated framework, it still makes mistakes. Individual misclassifications can occur, though large-scale trends remain informative.

- **Appropriate use cases** – The tool is designed primarily for scalable analyses of new subfields (where small errors cancel out in aggregate analyses) and *writing assistance* (offering feedback for authors on the writing, akin to systems such as SWIF^{2}T (Chamoun et al., 2024)). It is **not intended for peer review or paper acceptance decisions**, as it is not precise enough to make definitive judgments about research quality. Furthermore, our aim is **not** to prescribe how research should be framed, but to surface and clarify existing framings—whether practical, exploratory, or societal—to support reflection on how NLP contributions are positioned and interpreted.

- **Human oversight** – Automation enhances scalability but does **not** replace expert judgment. The tool should be used as a *supplementary resource*, requiring human review in critical applications.

Despite these limitations, our approach provides valuable insights into NLP research trends and can serve as a promising writing assistant, supporting both large-scale analysis and research framing refinement.

Ethical considerations

Our study analyzes publicly available NLP research papers to examine how they frame the intended use of their artefacts. All data used, including annotated excerpts, is drawn from these publications and included in our supplementary materials. Since we focus on textual analysis of already public documents, we do not anticipate ethical concerns. However, if any author requests that specific excerpts from their work be removed from our dataset or repository, we will comply immediately.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 173–184, New York, NY, USA. Association for Computing Machinery.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. 1984. Classification and regression trees. *Biometrics*, 40:874.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Eric Chamoun, Marzieh Saeidi, and Andreas Vlachos. 2023. Automated fact-checking in dialogue: Are specialized models needed? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16009–16020, Singapore. Association for Computational Linguistics.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong

Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, and Alex Kaskasoli. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv e-prints*, arXiv:2312.11805.

Sian Gooding, Lucia Lopez-Rivilla, and Edward Grefenstette. 2025. Writing as a testbed for open ended agents. *Preprint*, arXiv:2503.19711.

F. S. Grodzinsky, K. Miller, and M. J. Wolf. 2012. Moral responsibility for computing artifacts: "the rules" and issues of trust. *SIGCAS Comput. Soc.*, 42(2):15–25.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. *12th International Conference on Learning Representations, ICLR 2024, Messe Wien Exhibition and Congress Center, Vienna Austria May 7th, 2024 to May 11th, 2024*, 2024.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2).

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, and Adam Trischler. 2023. Responsible AI considerations in text summarization research: A review of current practices. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6246–6261, Singapore. Association for Computational Linguistics.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

A List of Epistemic Elements and Research Framings

For clarity and completeness, we list below the full set of canonical labels used in epistemic element extraction and research framing classification. For AFC, these labels are drawn from Schlichtkrull et al. (2023); full definitions can be found in their original paper. For HS, we introduce a new set of canonical labels, defined specifically for this study. Detailed descriptions of each label are provided in Appendix E.

A.1 Fact-checking

Epistemic Elements

- **Data Subjects**:
  professional journalists, citizen journalists, social media users, technical writers, public figures/politicians, product reviewers, suspects

- **Data Actors**:
  publishers, professional journalists, product reviewers, citizen journalists, scientists, media consumers, technical writers, engineers and curators, law enforcement, algorithm

- **Model Owners**:
  social media companies, law enforcement, scientists, professional journalists

- **Modeling Means**:
  classify/score veracity, classify/score stance, evidence retrieval, justification/explanation production, corpora analysis, human in the loop, generate claims, triage claims

- **Application Means**:
  identify claims, triage claims, supplant human fact-checkers, gather and present evidence, identify multimodal inconsistencies, automated removal, provide labels/veracity scores, provide aggregates of social media comments, filter system outputs, maintain consistency with knowledge base, analyse data, produce misinformation, vague persuasion

- **Epistemic Ends**:
  limit misinformation, limit ai-generated misinformation, increase veracity of published content, develop knowledge of NLP/language, avoid biases of human fact-checkers, detect falsehood for law enforcement

Research Framings

- *adversarial research, assisted external fact-checking, assisted internal fact-checking, assisted knowledge curation, assisted media consumption, automated content moderation, automated external fact-checking, scientific curiosity, truth-telling for law enforcement, vague debunking, vague identification, vague opposition*

A.2 Hate Speech Detection

Epistemic Elements

- **Data Subjects**:
  journalists, social media users, public figures/politicians, product reviewers, moderators, media consumers

- **Data Actors**:
  social media moderators, social media companies, scientists, media consumers, engineers and curators, law enforcement, algorithm

- **Model Owners**:
  media companies, social media companies, law enforcement

- **Modeling Means**:
  classify/score posts/text, provide justifications, human in the loop, corpora analysis, data collection

- **Application Means**:
  identify toxic content, supplant human moderators, gather and present justification, automated removal, provide labels/toxicity scores, filter system outputs, analyse data, vague persuasion, generate counter narratives

- **Epistemic Ends**:
  fight hate, counter hate, limit ai-generated toxicity/misinformation, develop knowledge of NLP/language, avoid biases of human moderators, detect hate/falsehood for law enforcement

Research Framings

- *assisted content moderation, assisted knowledge curation, automatic content moderation, LLM safety, scientific curiosity, truth-telling for law enforcement, vague data analysis, vague identification, vague moderation, vague opposition*

B Epistemic element extraction

B.1 Semantic clustering via NLI validation

Samples	Automated Fact-Checking						Hate Speech					
	Overall	DS	DA	MO	MM	EN	Overall	DS	DA	MO	MM	EN
1	68.18	82.12	68.32	96.90	52.25	67.81	69.59	70.38	67.04	96.82	59.24	69.19
3	70.17	83.21	71.11	96.90	54.76	71.44	72.34	74.12	71.82	97.13	60.09	72.05
5	71.59	84.29	72.65	97.02	**56.61**	73.93	72.52	74.68	71.84	96.82	60.84	73.11
8	72.12	84.73	73.97	97.45	56.29	73.50	73.80	73.78	72.74	**97.77**	62.09	**74.76**
10	**73.76**	**86.19**	**77.44**	**97.77**	56.21	**78.96**	**76.76**	**77.74**	**76.13**	96.60	**69.84**	74.04

Table 6: Filtered MRR (%) for each epistemic element across sample sizes. DS = Data Subjects, DA = Data Actors, MO = Model Owners, MM = Model Means, AM = Application Means, EN = Epistemic Ends.

C Research Framing Ranking

C.1 Logical mappings

C.1.1 Automated fact-checking

Adversarial Research:

$$\text{Data actors: } \textit{scientists}$$
$$\wedge \text{ Model means: } \textit{generate claims}$$
$$\wedge \text{ Application means: } \textit{produce misinformation}$$

Assisted External Fact-checking:

$$\bigvee \{\text{Data subjects: All except } \textit{professional journalists}, \textit{citizen journalists}, \textit{technical writers}\}$$
$$\wedge \bigvee \{\text{Data actors: } \textit{professional journalists}, \textit{citizen journalists}, \textit{technical writers}\}$$
$$\wedge \bigvee \{\text{Application means: All except } \textit{supplant human fact-checkers}, \textit{not specified}\}$$
$$\wedge \bigvee \{\text{Ends: } \textit{limit misinformation}, \textit{limit AI-generated misinformation}\}$$

Assisted Internal Fact-checking:

$$\bigvee \{\text{Data subjects: } \textit{professional journalists}, \textit{citizen journalists}, \textit{technical writers}\}$$
$$\wedge \bigvee \{\text{Data actors: } \textit{professional journalists}, \textit{citizen journalists}, \textit{technical writers}\}$$
$$\wedge \bigvee \{\text{Application means: All except } \textit{supplant human fact-checkers}, \textit{not specified}\}$$
$$\wedge \bigvee \{\text{Ends: } \textit{increase veracity of published content}, \textit{limit misinformation}\}$$

Assisted Knowledge Curation:

$$\bigvee \{\text{Data actors: } \textit{engineers and curators}, \text{Application means: } \textit{filter system outputs}\}$$
$$\wedge \bigvee \{\text{Model means: } \textit{classify/score veracity}, \textit{evidence retrieval}, \textit{human in the loop}\}$$
$$\wedge \text{Ends: } \textit{increase veracity of published content}$$

Assisted Media Consumption:

$$\bigvee \begin{cases} \text{(Data subjects: } \textit{social media users} \wedge \text{Model owners: } \textit{social media companies} \\ \quad \wedge \bigvee \{\text{Application means: } \textit{vague persuasion}, \text{ Ends: } \textit{limit misinformation}\}), \\ \text{Data actors: } \textit{media consumers} \end{cases}$$

Automated Content Moderation:

$$\bigvee \begin{cases} \text{(Data subjects: } \textit{social media users} \wedge \text{Data actors: } \textit{algorithm} \\ \quad \wedge \text{Model owners: } \textit{social media companies}), \\ \text{Application means: } \textit{automated removal} \end{cases}$$

Automated External Fact-checking:

$$\text{Application means: } \textit{supplant human fact-checkers}$$

Scientific Curiosity:

$$\left(\text{Data actors: } \textit{scientists} \wedge \text{Model owners: } \textit{scientists}\right)$$
$$\vee \text{ Ends: } \textit{develop knowledge of NLP/language}$$

Vague Debunking:

Data actors: *not specified*

$\wedge \bigvee \{\text{Model means: } \textit{evidence retrieval}, \textit{classify/score veracity}\}$

$\wedge \bigvee \{\text{Application means: } \textit{gather and present evidence}, \textit{provide labels/veracity scores}\}$

Vague Identification:

$$\text{Data actors: } \textit{not specified}$$
$$\wedge \text{ Application means: } \textit{identify claims}$$

Vague Opposition:

Application means: *not specified*

$\wedge \bigvee \{\text{Ends: } \textit{limit misinformation}, \textit{limit AI-generated misinformation}\}$

C.1.2 Hate Speech Detection

Assisted Content Moderation:

$$\bigvee \left\{ \begin{array}{l} \text{Model means: } \textit{human in the loop,} \\ \text{Data actors: } \textit{social media moderators} \end{array} \right\}$$

Assisted Knowledge Curation:

$$\text{Application means: } \textit{filter system outputs}$$
$$\wedge \text{ Data actors: } \textit{engineers and curators}$$

Automatic Content Moderation:

$$\bigvee \left\{ \begin{array}{l} \text{Application means: } \textit{supplant human moderators,} \\ \text{Application means: } \textit{automated removal,} \\ \text{Application means: } \textit{generate counter narratives} \end{array} \right\}$$

LLM Safety:

$$\text{Ends: } \textit{limit AI-generated toxicity/misinformation}$$

Scientific Curiosity:

$$\bigvee \left\{ \begin{array}{l} \text{Data actors: } \textit{scientists,} \\ \text{Application means: } \textit{analyze data,} \\ \text{Model means: } \textit{corpora analysis} \end{array} \right\} \wedge \text{ Ends: } \textit{develop knowledge of NLP/language}$$

Truth-telling for Law Enforcement:

$$\text{Data actors: } \textit{law enforcement}$$
$$\wedge \text{ Ends: } \textit{detect hate/falsehood for law enforcement}$$

Vague Data Analysis:

$$\bigvee \left\{ \begin{array}{l} \text{Application means: } \textit{analyze data,} \\ \text{Model means: } \textit{corpora analysis,} \\ \text{Model means: } \textit{data collection} \end{array} \right\}$$
$$\wedge \text{ Ends: } \textit{fight hate}$$

Vague Identification:

$$\text{Data actors: } \textit{not specified}$$
$$\wedge \text{ Application means: } \textit{identify toxic content}$$

Vague Moderation:

$$\bigvee \left\{ \begin{array}{l} \text{Data actors: } \textit{social media moderators,} \\ \text{Data actors: } \textit{social media companies,} \\ \text{Model owners: } \textit{social media companies} \end{array} \right\}$$
$$\wedge \text{ Data subjects: } \textit{social media users}$$
$$\wedge \bigvee \{\text{Application means: All except } \textit{supplant human moderators}, \textit{automated removal}\}$$
$$\wedge \bigvee \{\text{Model means: All except } \textit{human in the loop}\}$$

Vague Opposition:

Application means: *not specified*

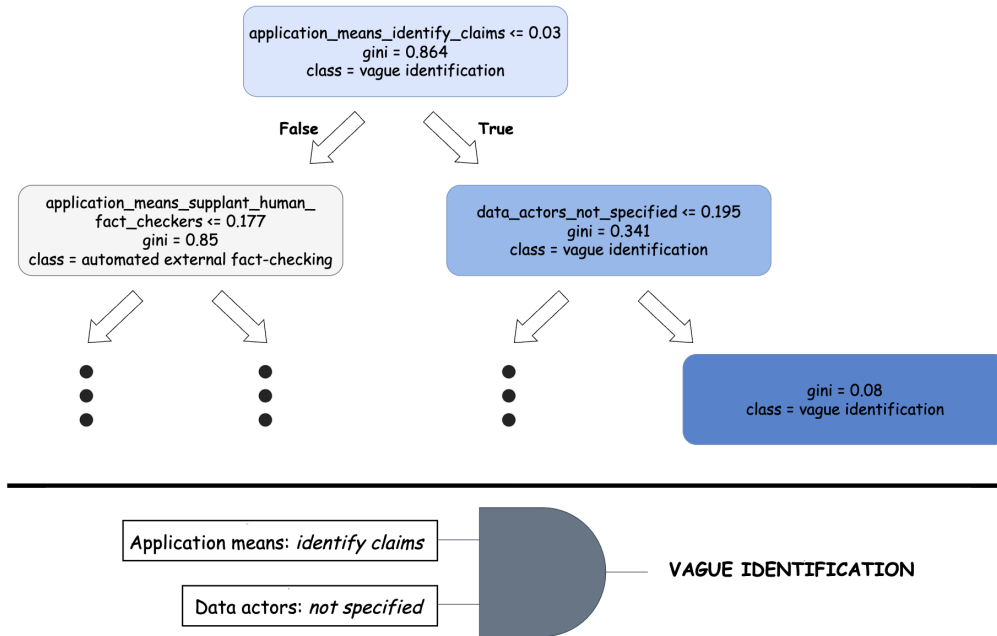∧ Ends: *fight hate*

C.2 Logical rule derivation example

Figure 3: Example illustrating how the logical rule for *Vague Identification* is derived from the decision tree.

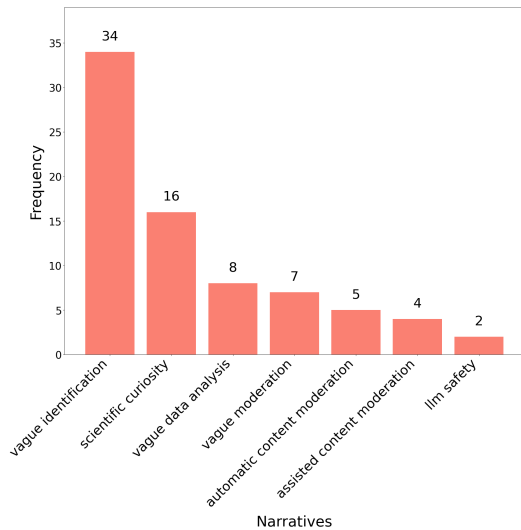## D Research Framings Distribution in Annotations

Hate Speech Dataset

Figure 4: Distribution of research framings identified by human annotators in our collected hate speech detection papers.
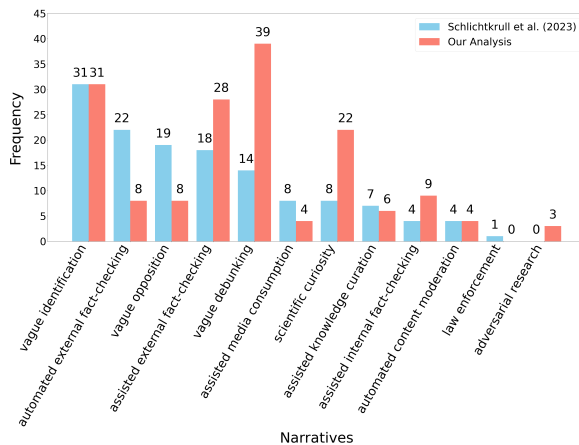
AFC Literature Analysis

Figure 5: Comparison of the distribution of research framings in our analysis (2023 onwards) and that of Schlichtkrull et al. (2023) (until 2023).

E Hate Speech Annotation Guidelines

To investigate the epistemology of hate speech detection and related tasks (e.g. bias in hate speech), we annotate the research framings of 49 research papers.

We adopt a two-step annotation process: (1) *paragraph-level annotation* to identify excerpts related to the paper's goals and methods, labeling *data subjects*, *data actors*, *model means*, *application means*, and *epistemic ends*; and (2) *discourse-level annotation* to infer the underlying research framings.

Our goal is to examine whether the means proposed in these research framings can realistically achieve the intended aims.

E.1 Paper Selection

We collect 49 research papers focused on hate speech detection and related tasks (e.g., bias in hate speech). Annotations are extracted from introductions and abstracts.

Initially, we annotated a small subset of papers based on predefined criteria. We refined category definitions through discussion, resolved label inconsistencies, and conducted a final annotation round to ensure consistency.

E.2 Annotation Criteria

We analyze each paper's introduction and abstract, structuring epistemic research framings at two levels: (1) *paragraph-level categorization* and (2) *research framing extraction*.

E.2.1 Paragraph-Level Annotation

At this level, we extract modeling and application means, epistemic goals, data actors, and data subjects. These elements are then combined into discourse-level research framings, e.g.:

"Social media moderators (actor) should use a classification model (modeling means) to triage posts (application means) from social media users (data subjects) to moderate online content (goal)."

Our annotation process follows an inductive approach: we begin with a predefined set of categories based on an initial pilot study. If, during annotation, we encounter means, ends, actors, or research framings that do not fit existing categories, we introduce new ones. We use a multi-label approach, as paragraphs and discourses may reference multiple categories.

E.2.2 Quotes

We extract epistemic quotes relevant to the paper's research framings, focusing on the *why* and *what*. Quotes are primarily sourced from the introduction; missing information is supplemented from the abstract.

E.2.3 Data Subjects

Entities whose texts are analyzed or moderated:

- Social media users

- Journalists

- Public figures/politicians

- Product reviewers

- Media consumers

Social Media Users   Includes anonymous contributors to social media, excluding explicitly public figures.

Journalists   Covers professional journalists and online publishers but excludes amateur journalists.

Public Figures/Politicians   Includes public figures such as politicians, actors, or influencers.

Product Reviewers   Individuals posting reviews on platforms like Amazon or Trustpilot.

Media consumers   End-users consuming content verified or enhanced by model outputs.

E.2.4   Data Actors
Entities intended to act on model outputs:

- Social media moderators

- Social media companies

- Scientists

- Media consumers

- Engineers and curators

- Law enforcement

- Algorithm

Social Media Moderators   Individuals employed to moderate social media content. Applicable only when explicitly stated that humans will act on model outputs.

Social Media Companies   Covers broader company-level decisions, including automated interventions based on model outputs.

Scientists   Includes researchers who analyze model outputs for scientific research rather than direct intervention.

Media Consumers   General users who encounter model outputs in content filtering systems.

Engineers and Curators   Covers those maintaining curated knowledge repositories such as Wikipedia.

Law Enforcement   Includes police, intelligence agents, or legal professionals utilizing model outputs.

Algorithm   Refers to fully automated systems that act on model outputs without human oversight.

E.2.5   Model Owners
Entities responsible for owning or deploying the models:

- Media companies

- Social media companies

- Law enforcement

Media Companies   Includes journalists and fact-checkers working for professional organizations.

Social Media Companies   Covers engineers and executives responsible for deploying moderation systems.

Law Enforcement   Includes agencies using model outputs for investigative or judicial purposes.

E.2.6   Modeling (ML) Means
The technical approach proposed in the paper:

- Classify/score posts or text

- Provide justifications

- Corpora analysis

- Data collection

- Human-in-the-loop

Classify/Score Posts or Text   When the paper proposes classifying posts (e.g., hate vs. non-hate, toxic vs. non-toxic) or using classifier-generated scores.

Provide Justifications   When the paper proposes generating explanations or justifications to support its goal (*epistemic ends*).

Corpora Analysis   When the methodology involves analyzing datasets or performing large-scale linguistic analysis.

Data Collection   When the paper focuses on collecting new datasets to address a specific problem.

Human-in-the-Loop   When human actors are integrated into the model's workflow, either as decision-makers or contributors.

E.2.7   Application Means
How the model is applied:

- Identify toxic content

- Supplant human moderators

- Gather and present justifications

- Automated removal

- Provide labels or toxicity scores

- Filter system outputs

- Analyze data

- Vague persuasion

- Generate counter narratives

Identify Toxic Content   ML models should identify harmful or offensive content from large volumes of online posts.

Supplant Human Moderators   Fully or partially replace human moderators by automating content moderation. If a human-in-the-loop approach is intended, this label is not applied.

Gather and Present Justifications   ML models should generate and present relevant justifications for why a post is harmful.

Automated Removal   ML models should autonomously remove posts from social media or other platforms without human intervention.

Provide Labels or Toxicity Scores   ML models should assign toxicity labels or scores to content. The target audience could be either users or platform administrators.

Filter System Outputs   ML models should act as filters to ensure that outputs from other models (e.g., LLMs) are free from toxicity or misinformation.

Analyze Data   Models and datasets should be used for research purposes, such as understanding the spread of hate speech or toxicity.

Vague Persuasion   ML models should influence users' perspectives through mechanisms such as warning labels or counter-narratives, though the exact approach is unspecified.

Generate counter narratives   : If the excerpt explicitly mentions that the released artefacts are designed to produce responses or statements aimed at refuting, mitigating, or opposing harmful or hateful content (e.g., generating automated replies to hate speech).

E.2.8   Epistemic Ends
The overarching goal of the research:

- Fight hate

- Counter hate

- Limit AI-generated toxicity

- Develop knowledge of NLP/language

- Avoid biases of human moderators

- Detect hate for law enforcement

Fight Hate   The ultimate aim is to reduce the spread or influence of hate speech.

Counter Hate   The goal is to challenge stereotypes and harmful narratives.

Limit AI-Generated Toxicity   The objective is to mitigate the risks of AI-generated harmful or toxic content.

Develop Knowledge of NLP/Language   Research is conducted to better understand language structures and improve NLP models.

Avoid Biases of Human Moderators   The aim is to develop moderation systems that overcome human biases.

Detect Hate for Law Enforcement   The model is intended for use in legal or law enforcement contexts, such as aiding police investigations or courtroom decisions.

E.2.9   Important Note
We rely strictly on explicit statements in extracted quotes, avoiding inferred or implied meanings.

Research Framings	Ordinal Class
Vague Opposition	-3
Vague Identification	-2
Vague Moderation	-1
Law Enforcement	0
Scientific Curiosity	1
Automatic Content Moderation	2
Assisted Content Moderation	3
Assisted Knowledge Curation	4

Table 7: Ordinal classes for research framings: underspecified research framings have negative values.

E.3 Research Framings

At the discourse level, we extract the epistemic research framings present in the paper. We envision research framings as discourse structures combining the (modeling/application) means, ends, data actors, and data subjects extracted at the paragraph level.

LLM Safety The paper mentions limiting toxic publishes or generated content as an end.

Vague Identification The paper mentions identifying or detecting toxic posts as the means, and fighting hate as the end. However, it is not clear how the authors intend to accomplish that end using proposed means. Typically, there are no data actors – it is also not clear who should act on the model predictions.

Vague Moderation When the paper proposes to assist the moderation process without explaining how. It is clear that the mechanism is supposed to follow what moderators are currently doing, but not where or how the ML model will be used in this process. Furthermore, it is *unclear* whether the entire process will be automated. For example, the paper may suggest that a model should be used by moderators, but it is not clear *how* the model assists in that. A common warning sign is a mismatch between ML means and application means: classification is not really useful for moderators (or social media users), and so papers suggesting to use those either to assist or automate moderation are often not clear in how.

Vague Opposition Restricted to cases without any application means (model means and no application means in contrast to vague identification and vague debunking). E.g. *a machine learning/automated system will reduce the spread of misinfor-*

mation.
When the paper presents a research framing of vague opposition to misinformation. The *end* is to limit the spread or influence of misinformation, and the *ML means* are, for example, to classify claims. However, the connection between means and ends is left unmentioned, and epistemic actors are typically entirely absent. An impression is given that the development of automated fact-checking will limit the spread of misinformation, but the link between the two is left unstated.

Assisted Content Moderation If the paper proposes the deployment of automated moderation as a tool to assist content moderators on social media platforms. Here, the *end* is to fight hate on social media platforms, the *means* are to provide suggestions for posts to delete (along with, potential justifications/explanations), and the *actors* are human moderators who make the final choice on whether posts should be deleted or not.

Automatic Content Moderation If the paper analysed proposes a similar content moderation strategy, but instead of assisting human moderator, it suggests replacing them entirely. In this case, the *end* is to fight hate online, the *means* are to deploy classifiers to flag posts and remove any labelled offensive, and the *actors* are the executives and engineers at social media companies who deploy and make decisions about such systems.

Assisted Knowledge Curation When the analysed paper proposes toxic content detection primarily as a component filtering the information kept in some curated knowledge vault, including graph-based knowledge bases as well as text-based collections such as Wikipedia. Here, the *end* is to avoid toxicity/stereotypical content of the knowledge vault, the *means* are to use automated fact-checking as an additional toxic content filter layer that prevents disputed content from being added (or interrogates already added content), and the *actors* are typically the engineers who maintain the knowledge base.

Scientific Curiosity When the authors of the analysed paper justified their projects purely on the basis of scientific curiosity. While differing strongly from the other research framings presented here, this is still a virtue epistemological research framing, concerned with the production of *good knowledge*. Here, the *end* is to increase scientific knowledge of semantics; the *means* are to

learn how to build automated systems that mimic human moderators, a process theorised to yield knowledge about the construction of meaning; and the *actors* are scientists in natural language processing and adjacent fields.

Vague Data Analysis   When the authors of the analysed paper justify their data analysis on the basis of scientific curiosity but also claim that the end goal is to fight hate without a clear explanation about how to achieve this.

F   Prompts

F.1   Epistemic Element Extraction: Hate speech detection

You will be provided with the introduction of a computer science research paper on hate speech detection and related tasks to understand the broader context of the study. Afterward, you will be given a specific excerpt from the paper. Your task is to answer the question based solely on the specific excerpt while considering the broader context provided in the introduction.

Introduction: {introduction}
Specific Excerpt: {excerpt}

F.1.1   Data Subjects

Task: Identify the **data subjects** explicitly mentioned by the authors in the provided excerpt. Data subjects are individuals or entities that the authors explicitly identify as the intended targets of the artefacts released in this study. Your answer should be based solely on the provided excerpt.

Key Rules:

1. **Explicit mention only**: Consider only data subjects explicitly named or described in the excerpt.
2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose.
3. **Default to 'not specified in this paragraph'**: If no data subjects are explicitly mentioned, respond accordingly.
4. **Consider the big picture**: Identify who the authors envision as targets for their released artefacts.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned data subjects?

You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Journalists**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from journalists, including online publishers.
- **Social media users**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from contributors to social media platforms, excluding public figures. This includes commenters on forums as well as, e.g., Twitter users. Similarly, editors of Wikipedia and other collaborative writing projects fall under this category as well.
- **Public figures/politicians**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from public figures like politicians or actors. These statements could be from media releases, interviews, speeches, or similar.
- **Product reviewers**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from individuals reviewing products on platforms like Amazon or Trustpilot.
- **Media consumers**: If the excerpt explicitly mentions that the released artefacts are expected to be applied on statements by end-users consuming content verified or enhanced by model outputs.

If none of these labels apply, respond with "not specified in this paragraph."

F.1.2   Data Actors

Task: Identify the **data actors** explicitly mentioned in the provided excerpt.

Data actors are individuals or entities expected to act on or utilize the artefacts released in this study.

Key Rules:

1. **Explicit mention only**: Consider only explicitly named data actors.
2. **Focus on the released artefacts**: Ensure relevance to the specific artefacts described in the excerpt.
3. **Default to 'not specified in this paragraph'**: If no data actors are explicitly mentioned, respond accordingly.
4. **Consider the big picture**: Identify who the authors expect to act on the model outputs.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned data actors?

Identify the individuals or entities expected to act on the model outputs in the excerpt.

Possible Labels:

- **Social media moderators**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by individuals hired to moderate social media spaces, i.e., explicitly stated that a human employee should act on model outputs.
- **Social media companies**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by social media companies, including engineers working to maintain the network. This is applicable when decisions should be made based on model outputs or when it is unclear.
- **Scientists**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by researchers using model outputs for scientific analysis or experiments. E.g., to analyze data with the express purpose of learning something about it, not acting on the model decisions.

- **Media consumers**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by end-users consuming content verified or enhanced by model outputs. Only applicable when the consumer is directly understood (possibly implicitly) to use the system, e.g., in the form of a browser extension. The decision to use the outputs of the tool must be in the hands of the consumer.
- **Engineers and curators**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by maintainers of knowledge bases integrating model outputs.
- **Law enforcement**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by agents using model outputs in investigations or legal processes.
- **Algorithm**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by automated systems acting directly on model outputs without human intervention. E.g., remove posts based on the model's predictions.

If none of these labels apply, respond with "not specified in this paragraph."

F.1.3 Model Owners

Task: Identify the **model owners** explicitly mentioned by the authors in the provided excerpt. Model owners are entities responsible for controlling or maintaining the artefacts released in this study.

Key Rules:

1. **Explicit mention only**: Consider only model owners explicitly named in the excerpt.
2. **Focus on the released artefacts**: Do not assume ownership unless explicitly stated.
3. **Default to 'not specified in this paragraph'**: If no entities are explicitly mentioned as owners, respond accordingly.

4. **Consider the big picture**: Identify the responsible institutions or organizations.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned model owners?

Determine the entities responsible for maintaining the model.

Possible Labels:

- **Media companies**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by journalists and hate speech detection entities working for professional (non-amateur) media companies.
- **Social media companies**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by organizations maintaining social platforms and using models for moderation, including the engineers working to maintain the social network.
- **Law enforcement**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by agencies employing models for investigations and truth verification.

If none of these labels apply, respond with "not specified in this paragraph."

F.1.4 Modeling (ML) Means

Task: Identify the **modeling means** explicitly mentioned in the provided excerpt. Modeling means refer to the specific machine learning methods or strategies proposed by the authors.

Key Rules:

1. **Explicit mention only**: Consider only modeling methods explicitly named in the excerpt.
2. **Focus on the released artefacts**: Do not infer methods unless explicitly stated.

3. **Default to 'not specified in this paragraph'**: If no methods are explicitly mentioned, respond accordingly.
4. **Consider the big picture**: Identify the modeling strategies used in the study.

Question: Based on the provided excerpt and using the provided labels, what are the explicitly mentioned modeling means?

Identify the machine learning methods proposed in the excerpt.

Possible Labels:

- **Classify/score posts/text**: If the excerpt explicitly mentions that the released artefacts involve classifying posts (hate/toxic and non-hate/non-toxic etc.) or using the scores returned by the classifier.
- **Provide justifications**: If the excerpt explicitly mentions that released artefacts involve generating explanations for model decisions to reach its goals (ends).
- **Human in the loop**: If the excerpt explicitly mentions that released artefacts involve including human oversight in the solution or main process described in the paper.
- **Corpora analysis**: If the excerpt explicitly mentions that released artefacts involve using data analytics or analyzing datasets for patterns or insights.
- **Data collection**: If the excerpt explicitly mentions that released artefacts involve collecting data to solve a given problem.

If none of these labels apply, respond with "not specified in this paragraph."

F.1.5 Application Means

Task: Identify the **application means** explicitly mentioned by the authors in the provided excerpt. Application means refer to the concrete, actionable uses of the artefacts released in the study.

Key Rules:

1. **Explicit mention only**: Consider only application means explicitly described in the excerpt.
2. **Focus on the released artefacts**: Differentiate between general discussion and specific proposals.
3. **Default to 'not specified in this paragraph'**: If no application means are mentioned, respond accordingly.
4. **Consider the big picture**: Identify how the authors envision their artefacts being applied.

Question: Based on the provided excerpt and using the provided labels, what are the explicitly mentioned application means?

Identify the concrete application of the model.

Possible Labels:

- **Identify toxic content**: If the excerpt explicitly mentions that released artefacts would be leveraged to detect potentially harmful/offensive posts on the Internet.
- **Supplant human moderators**: If the excerpt explicitly mentions that released artefacts would be leveraged to replace human moderators entirely or at least partially by automatically and independently handling some content in real-world scenarios. If the intention is a human-in-the-loop system, supplanting human moderators is not the means.
- **Gather and present justification**: If the excerpt explicitly mentions that released artefacts would be leveraged to find relevant justification for why a post is harmful, with the aim of showing it to a human.
- **Automated removal**: If the excerpt explicitly mentions that released artefacts would be leveraged to automatically delete or hide claims identified as false, harmful, or in violation of platform policies.
- **Provide labels/toxicity scores**: If the excerpt explicitly mentions that released artefacts would be leveraged to assign toxicity scores, labels, or indicators to individual claims and show them to users/human annotators.
- **Filter system outputs**: If the excerpt explicitly mentions that released artefacts would be leveraged to refine or validate the outputs of other machine learning systems to improve accuracy or alignment.
- **Analyze data**: If the excerpt explicitly mentions that released artefacts would be leveraged to examine and study patterns or trends in datasets of content related to toxicity or hate speech detection to get a better understanding of how hate/toxicity spreads.
- **Vague persuasion**: If the excerpt explicitly mentions that released artefacts would be leveraged to influence user decisions or beliefs (e.g., through warning labels) without specifying the exact method of persuasion.
- **Generate counter-narratives**: If the excerpt explicitly mentions that the released artefacts are designed to produce responses or statements aimed at refuting, mitigating, or opposing harmful or hateful content (e.g., generating automated replies to hate speech).

If none of these labels apply, respond with "not specified in this paragraph."

F.1.6 Ends

Task:

Identify the **ends** explicitly mentioned by the authors in the provided excerpt. The ends represent the ultimate goals or outcomes of the released artefacts.

Key Rules:

1. **Explicit mention only**: Consider only explicitly mentioned goals or outcomes.

2. **Focus on the released artefacts**: Do not assume end goals unless explicitly stated.

3. **Default to 'not specified in this paragraph'**: If no end goals are mentioned, respond accordingly.

4. **Consider the big picture**: Identify the broader societal or epistemic impact.

Question: Based on the provided excerpt and using the provided labels, what are the explicitly mentioned ends? Determine the ultimate goal of the study.

Possible Labels:

- **Fight hate**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to prevent hate from spreading or to limit its influence.

- **Counter hate**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to counter stereotypes and harmful narratives.

- **Limit AI-generated toxicity/misinformation**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to prevent AI-generated harmful/toxic content from spreading.

- **Develop knowledge of NLP/language**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to advance understanding of language and NLP methods. Authors would be studying the difficult problem of hate speech detection to learn more about how language works and how to build models that interpret semantics.

- **Avoid biases of human moderators**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to achieve unbiased and consistent hate speech detection.

- **Detect hate/falsehood for law enforcement**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to enable law enforcement to identify and act on hate speech.

If none of these labels apply, respond with "not specified in this paragraph."

F.2 Epistemic Element Extraction: Automated Fact-checking

You will be provided with the introduction of a computer science research paper on automated fact-checking to understand the broader context of the study. Afterward, you will be given a specific excerpt from the paper. Your task is to answer the question based solely on the specific excerpt while considering the broader context provided in the introduction.

Introduction: {introduction}

Specific Excerpt: {excerpt}

F.2.1 Data Subjects

Task: Identify the **data subjects** explicitly mentioned by the authors in the provided excerpt. Data subjects are the individuals or entities that the authors explicitly identify as the intended targets of the artefacts released in this study. The possible data subjects will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only data subjects explicitly named or described in the excerpt. Do not infer their presence unless they are directly identified as targeted by the released artefacts.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose. If the authors discuss fact-checking broadly or reference prior work, do not assume relevance; focus solely on the data subjects tied to the artefacts they release. For example, if the text discusses fake news or its impact, do not assume social media users are data subjects unless explicitly mentioned as being the intended targets. Similarly, if politics is mentioned, do not assume politicians are data subjects unless they are explicitly identified.

3. **Default to 'not specified in this paragraph'**: If no data subjects are explicitly mentioned or described as targeted to use the released artefacts, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: The goal is to automatically identify who the authors envision as targets for their released automated fact-checking artefacts, if mentioned at all. Consider the study's broader purpose and how it is represented in the specific excerpt.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned data subjects?

You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Professional Journalists**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from professional journalists, including fact-checkers.

- **Citizen Journalists**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from amateur journalists or collectives working without formal training.

- **Social Media Users**: If the excerpt explicitly mentions that the released artefacts are de-signed to be applied on statements from contributors to social media platforms, excluding public figures.

- **Technical Writers**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from writers of technical or scientific documents, including legal professionals, to spot errors in their articles before publication.

- **Public Figures/Politicians**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from public figures like politicians or actors.

- **Product Reviewers**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from individuals reviewing products on platforms like Amazon or Trustpilot.

- **Suspects**: If the excerpt explicitly mentions that the released artefacts are designed to be applied on statements from people under investigation or scrutiny, often in a legal or forensic context.

If none of these labels apply, respond with "not specified in this paragraph."

F.2.2 Data Actors

Task: Identify the **data actors** explicitly mentioned in the provided excerpt. Data actors are individuals or entities that the authors specifically identify as expected to act on or utilize the artefacts released in this study. The possible data actors will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only data actors explicitly named or described in the excerpt. Do not infer their presence unless they are directly identified as acting on the released artefacts.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose. If the authors discuss fact-checking broadly or reference prior work, do not assume relevance; focus solely on the data actors tied to the artefacts they release. For example, if the text discusses fake news or its impact, do not assume journalists, moderators, or consumers are data actors unless explicitly mentioned as acting on the model's outputs. Similarly, if a suspect or individual is mentioned, do not assume they are a data actor unless explicitly described as targeted to interact with the model.

3. **Default to 'not specified in this paragraph'**: If no data actors are explicitly mentioned or described as targeted to use the released artefacts, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: The goal is to automatically identify who the authors envision as actors for their released automated fact-checking artefacts, if mentioned at all.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned data actors?

You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Publishers**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by organizations or individuals that distribute printed or online content, such as books and comics.

- **Professional journalists**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by professional journalists, including fact-checkers and reporters to get model insights on their work.

- **Product reviewers**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by reviewers providing insights or opinions about products.

- **Citizen journalists**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by amateur reporters leveraging model outputs for investigative purposes.

- **Scientists**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by researchers using model outputs for scientific analysis or experiments.

- **Media consumers**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by end-users consuming content verified or enhanced by model outputs.

- **Technical writers**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by writers improving technical or legal documents with model suggestions.

- **Engineers and curators**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by maintainers of knowledge bases integrating model outputs.

- **Law enforcement**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by agents using model outputs in investigations or legal processes.

- **Algorithm**: If the excerpt explicitly mentions that the released artefacts are expected to be utilized by automated systems acting directly on model outputs without human intervention.

If none of these labels apply, respond with "not specified in this paragraph."

F.2.3 Model Owners

Task: Identify the **model owners** explicitly mentioned by the authors in the provided excerpt. Model owners are entities or organisations that the authors explicitly target as being responsible for controlling or maintaining the artefacts released in this study. The possible model owners will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only model owners explicitly named or described in the excerpt. Do not infer their presence unless they are directly identified as expected to own the released artefacts.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose. If the authors discuss fact-checking broadly or reference prior work, do not assume relevance; focus solely on the model owners tied to the artefacts they release. For example, if the text mentions researchers conducting experiments, it does not imply they own the models unless explicitly stated.

3. **Default to 'not specified in this paragraph'**: If no entities or organisations are explicitly mentioned or intended to own the released artefacts, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: The goal is to automatically identify who the authors envision as owning their released automated fact-checking artefacts, if mentioned at all.

   Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned model owners?
The task follows the same structure as for Data Subjects, but with the following labels:

- **Social media companies**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by organizations maintaining social platforms and using models for moderation.

- **Law enforcement**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by agencies employing

models for investigations and truth verification.

- **Scientists**: If the excerpt explicitly mentions that the released artefacts are expected to be owned by researchers developing or owning models for analysis and experiments.

If none of these labels apply, respond with **"not specified in this paragraph"**.

F.2.4 Model Means

Task: Identify the **model means** explicitly mentioned by the authors in the provided excerpt. Model means are the specific machine learning methods or strategies proposed by the authors in their study. The possible model means will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only model means explicitly named or described in the excerpt. Do not infer their presence unless they are directly identified.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose. If the authors discuss fact-checking broadly or reference prior work, do not assume relevance; focus solely on the model means tied to the artefacts they release.

3. **Default to 'not specified in this paragraph'**: If no specific methods or strategies are mentioned in the excerpt, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: Identify the strategies used by the authors in their study, if mentioned at all.

   Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned model means?
You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Classify/score veracity**: If the excerpt explicitly mentions that the released artefacts involve determining if claims are true or false.

- **Classify/score stance**: If the excerpt explicitly mentions that the released artefacts involve evaluating whether evidence supports or opposes claims.

- **Evidence retrieval**: If the excerpt explicitly mentions that released artefacts involve locating relevant evidence to support or refute claims.

- **Justification/explanation production**: If the excerpt explicitly mentions that released artefacts involve generating explanations for model decisions.

- **Corpora analysis**: If the excerpt explicitly mentions that released artefacts involve analyzing datasets for patterns or insights.

- **Human in the loop**: If the excerpt explicitly mentions that released artefacts involve including human oversight in the solution or main process described in the paper.

- **Generate claims**: If the excerpt explicitly mentions that released artefacts involve producing synthetic claims for adversarial training or testing.

- **Triage claims**: If the excerpt explicitly mentions that released artefacts involve ranking claims by importance for efficient processing.

If none of these labels apply, respond with "not specified in this paragraph."

F.2.5 Application Means

Task: Identify the **application means** explicitly mentioned by the authors in the provided excerpt. Application means refer to the concrete, actionable uses of the artefacts released in the study, as proposed by the authors, to directly support the study's ultimate goal. The possible application means will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only application means explicitly named or described in the excerpt.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose.

3. **Default to 'not specified in this paragraph'**: If no application means are mentioned in the excerpt, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: Identify how the authors envision their automated fact-checking artefacts to be applied in a real-world scenario.

You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Identify claims**: If the excerpt explicitly mentions that released artefacts would be leveraged to detect potentially false or misleading statements.

- **Triage claims**: If the excerpt explicitly mentions that released artefacts would be leveraged to prioritize claims based on urgency or relevance.

- **Supplant human fact-checkers**: If the excerpt explicitly mentions that released artefacts would be leveraged to replace human fact-checkers entirely or partially.

- **Gather and present evidence**: If the excerpt explicitly mentions that released artefacts would be leveraged to retrieve and organize evidence supporting or refuting specific claims.

- **Identify multimodal inconsistencies**: If the excerpt explicitly mentions that released artefacts would be leveraged to detect contradictions between different content types.

- **Automated removal**: If the excerpt explicitly mentions that released artefacts would be leveraged to automatically delete or hide claims identified as false or harmful.

- **Provide labels/veracity scores**: If the excerpt explicitly mentions that released artefacts would be leveraged to assign truthfulness scores or labels.

- **Provide aggregates of social media comments**: If the excerpt explicitly mentions that released artefacts would be leveraged to summarize trends or sentiments from social media comments.

25328

- **Filter system outputs**: If the excerpt explicitly mentions that released artefacts would be leveraged to refine the outputs of other machine learning systems.

- **Maintain consistency with knowledge base**: If the excerpt explicitly mentions that released artefacts would be leveraged to ensure outputs align with established facts.

- **Analyze data**: If the excerpt explicitly mentions that released artefacts would be leveraged to study patterns in misinformation datasets.

- **Produce misinformation**: If the excerpt explicitly mentions that released artefacts would be leveraged to generate false claims for research purposes.

- **Vague persuasion**: If the excerpt explicitly mentions that released artefacts would be leveraged to influence user decisions without specifying the method.

If none of these labels apply, respond with "not specified in this paragraph."

F.2.6 Ends

Task: Identify the **ends** explicitly mentioned by the authors in the provided excerpt. The ends represent the ultimate goals or outcomes of the released artefacts, reflecting how the authors envision their contributions to society. The possible ends will be outlined below, and your answer should solely be based on the provided excerpt.

Key rules:

1. **Explicit mention only**: Consider only ends explicitly named or described in the excerpt.

2. **Focus on the released artefacts**: Differentiate between general background information and what the authors specifically propose.

3. **Default to 'not specified in this paragraph'**: If no ends are mentioned in the excerpt, respond with **"not specified in this paragraph"**.

4. **Consider the big picture**: Identify how the authors envision the impact of their released automated fact-checking artefacts in the real world.

Question: Based on the provided excerpt and using the provided labels, who are the explicitly mentioned ends? You will select the single most appropriate label that answers the question. Provide reasoning then choose only from the following list:

- **Limit misinformation**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to prevent misinformation from spreading.

- **Limit AI-generated misinformation**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to prevent AI-generated falsehoods.

- **Increase veracity of published content**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to ensure accuracy of published material.

- **Develop knowledge of NLP/language**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to advance understanding of language and NLP methods.

- **Avoid biases of human fact-checkers**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to achieve unbiased and consistent fact-checking.

- **Detect falsehood for law enforcement**: If the excerpt explicitly mentions that the ultimate goal of the released artefacts would be to enable law enforcement to identify and act on falsehoods.

If none of these labels apply, respond with "not specified in this paragraph."

F.3 Research narrative classification: Hate Speech Detection

You will be provided with the introduction of a computer science research paper on hate speech detection. Your task is to identify the research narratives that describe the intended uses of the artefacts presented in the study. You will be provided with the introduction of the study.

Introduction:
{introduction}
Task: Your task is to select the research narratives that best describe the intended uses of the artefacts (e.g., datasets, models) introduced in the paper.

- **Reasoning:** Provide an explanation of why each selected narrative(s) apply based on the introduction.

- **Selected Labels:** Choose the most appropriate label(s) from the list below. If more than one applies, separate them with commas.

  Available research narratives:

F.3.1 LLM Safety

Definition: The paper discusses limiting toxic content generated by Large Language Models (LLMs). The system is designed to filter or prevent the publishing of toxic or harmful content.

Example:

LLMs can generate unsafe content, which harms users. We design a hate speech classifier to serve as a filter on the output of our LLM.

Mapping: This narrative typically involves the end goal of limiting AI-generated toxicity or misinformation.

F.3.2 Vague Identification

Definition: The paper mentions identifying or detecting toxic posts as the means, and fighting hate as the end. However, it is unclear how the authors intend to accomplish that end using those means.

Example:

Hate speech is a significant issue on social media, leading to real-world harm. It is therefore crucial to develop models capable of identifying and detecting toxic content.

Mapping: This narrative is identified when the application means focus on identifying toxic content, but data actors are not specified.

F.3.3 Vague Moderation

Definition: The paper proposes to assist the moderation process but does not explain how. The connection between the proposed system and how it assists moderators remains unclear.

Example:

Models will be used for debunking (but not how).

Mapping: This narrative typically appears when the paper mentions assisting moderators but does not specify exactly how the assistance works. E.g. it is not specified whether a human will be involved in the moderation process or whether the tool is designed to automatically removing hate posts/supplant human moderators.

F.3.4 Vague Opposition

Definition: The paper presents a narrative of opposition to online hate speech without explicitly mentioning how this opposition will be enacted.

Example:

Hate speech is a major societal problem, eroding community trust and leading to real-world crimes. It is therefore of paramount importance that the spread of hate speech is stopped. Automated classification of hateful messages represents one solution to this critical problem.

Mapping: This narrative is identified when the application means are not specified, but the ends include fighting hate.

F.3.5 Assisted Content Moderation

Definition: The paper proposes the deployment of automated moderation as a tool to assist content moderators on social media platforms. The system provides suggestions for posts to delete, but the final decision is left to human moderators.

Example:

Social media is rife with hate speech, eroding community trust and harming users. One

solution is for moderators to remove harmful content. However, with the volume of posts made every day, this strategy is too costly. In this paper, we develop an automated system for filtering posts, helping moderators quickly discover and make decisions on circulating content.

Mapping: This narrative typically includes human moderators as data actors that are involved in the moderation process (no full automation of the process).

F.3.6 Automatic Content Moderation

Definition: The paper proposes replacing human content moderators with fully automated systems. The means involve deploying classifiers to automatically detect and remove flagged posts.

Example:

Social media is rife with online hate, leading to real-world consequences. A costly but effective solution is to remove flagged content. In this paper, we develop an automated system for detecting harmful posts, which serves as a first line of defense against online hate.

Mapping: This narrative typically includes application means focused on automated removal, supplanting human fact-checkers or generating counter narratives.

F.3.7 Assisted Knowledge Curation

Definition: The paper proposes toxic content detection primarily as a filtering component for curated knowledge vaults such as Wikipedia or knowledge graphs. The goal is to prevent the addition of harmful content.

Example:

Knowledge bases fuel many real-world NLP applications, such as question answering. The maintenance of knowledge bases is an expensive process, yet as knowledge evolves, they must be kept up-to-date. Automated hate speech detection systems could be used to prevent harmful content from being added or to flag existing content for review.

Mapping: This narrative typically includes engineers and curators as data actors, with application means focused on filtering system

outputs to maintain high-quality knowledge bases.

F.3.8 Scientific Curiosity

Definition: The paper justifies its research purely on the basis of scientific curiosity, focusing on developing better NLP models without a specific application goal.

Example:

Online moderation is a complex task requiring reasoning about disputed content. For systems to accurately detect harmful language, significant linguistic understanding is necessary. As such, automated moderation is an ideal testbed for developing and evaluating new NLP models.

Mapping: This narrative typically includes scientists as data actors or a certain analysis, with the primary end being to develop knowledge of NLP/language.

F.3.9 Vague Data Analysis

Definition: The paper justifies its data analysis as contributing to scientific knowledge while also claiming that the goal is to fight hate, but without a clear connection between analysis and action.

Example:

Hate speech harms social media users. It is therefore crucial to understand the semantics of common hateful phrases. We create and analyze a dataset to this end.

Mapping: This narrative typically includes application means focused on analyzing data or corpora analysis but lacks a clear connection between this analysis and real-world interventions.

F.3.10 Truth-Telling for Law Enforcement

Definition: The paper proposes hate speech detection primarily for use in law enforcement contexts, such as detecting potential threats or identifying individuals engaged in online hate speech.

Example:

Hate speech is illegal in some jurisdictions, but the police do not have time to monitor

all instances. We develop a classifier for social media posts that can be used to identify potential perpetrators.

Mapping: This narrative typically includes law enforcement as data actors, with the end goal of detecting hate speech for use in investigations.

Prediction of the ranking model:

    {Ranking   explanations   split   by
    confidence levels}

Using your judgment and reasoning, identify the narrative(s) in the provided excerpt.

F.4 Research narrative classification: Automated Fact-checking

You will be provided with the introduction of a computer science research paper on automated fact-checking. Your task is to identify the research narratives that describe the intended uses of the artefacts presented in the study. You will be provided with the introduction of the study.

Introduction:

    {introduction}

Task: Your task is to select the research narratives that best describe the intended uses of the artefacts (e.g., datasets, models) introduced in the paper.

- **Reasoning:** Provide an explanation of why each selected narrative(s) apply based on the introduction.

- **Selected Labels:** Choose the most appropriate label(s) from the list below. If more than one applies, separate them with commas.

  Available research narratives:

F.4.1 Vague identification

Definition: The paper mentions identifying or detecting misinformative claims as the means, and limiting misinformation as the end. However, it is not clear how the authors intend to accomplish that end using those means. Typically, there are no data actors – it is also not clear who should act on the model's predictions. Vague identification applies only in cases where the authors say that they want to identify or detect misinformation without saying what they are going to do with that afterward, e.g., in rumor detection papers where they claim that they want to detect rumors but

we do not know what they want to do with this identification or classification labels.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g., inducing hesitance to adopt life-saving vaccines. The amount of messages spread on social media has increased drastically in recent years. Therefore, it is necessary to automatically identify potentially misinformative claims in order to address this problem.

Mapping: This narrative is typically identified when the application means focus on identifying claims, but data actors are not specified.

F.4.2 Automated External Fact-Checking

Definition: When the paper proposes fully automated external fact-checking, including all parts of the pipeline. It is only automated external fact-checking when the authors explicitly say the process should be automated; otherwise, it is vague debunking.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g., inducing hesitance to adopt life-saving vaccines. An important way to fight misinformation is the production of relevant counter-messaging, i.e., the work done by organizations such as Full Fact or PolitiFact. With the number of false claims published on social media every hour, it is not feasible for human journalists to debunk them all. As such, the process must be automated.

Mapping: This narrative typically includes application means designed to supplant human fact-checkers, automating the entire fact-checking process.

F.4.3 Vague Opposition

Definition: Restricted to cases without any application means (model means and no application means in contrast to vague identification and vague debunking). E.g., a machine learning/automated system will reduce the spread of misinformation. When the paper presents a narrative of vague opposition to misinformation. The end is to limit the

spread or influence of misinformation. However, the connection between modeling means and ends is left unmentioned. An impression is given that the development of automated fact-checking will limit the spread of misinformation, but the link between the two is left unstated.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g., inducing hesitance to adopt life-saving vaccines. It is therefore of paramount importance that the spread of false information is stopped. Automated fact-checking – that is, the automatic classification of claim veracity – represents one solution to this critical problem.

Mapping: Application means are typically not specified, but the ends typically include limiting misinformation or limiting AI-generated misinformation.

F.4.4   Assisted External Fact-Checking

Definition: When the paper proposes to deploy automated fact-checking as an assistive tool for journalists, deployed for external fact-checking. This is restricted to when the paper proposes to improve one or multiple components in the automated fact-checking pipeline rather than the whole pipeline/end-to-end system (in the latter case it becomes automated external fact-checking). I.e., when it is human-in-the-loop, then it is assisted fact-checking but not automated.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g., inducing hesitance to adopt life-saving vaccines. An important way to fight misinformation is the production of relevant counter-messaging, i.e., the work done by organizations such as Full Fact or PolitiFact. With the number of false claims published on social media every hour, it is not feasible for human journalists to debunk them all. Journalists could use automated fact-checking to triage incoming claims to limit the workload, or to quickly surface relevant evidence while producing articles.

Mapping: This narrative typically involves data actors such as professional journalists, citizen journalists, or technical writers, supported by modeling means like human-in-the-loop, classify/score veracity, or evidence retrieval.

F.4.5   Vague Debunking

Definition: The authors propose to produce evidence-based debunkings of text via automated means, but it is not clear whether the introduced model is an assistive tool for fact-checkers or fully automated. Further, it is not clear how the "debunkings" will be communicated to misinformation-believers. It is clear that the mechanism is supposed to follow what fact-checkers are currently doing, but not where or how the ML model will be used in this process. Furthermore, it is unclear whether the entire process will be automated. For example, the paper may suggest that an automated fact-checking model should be used by fact-checkers, but it is not clear how the model assists in that.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g., inducing hesitance to adopt life-saving vaccines. One proposed solution is to debunk circulating claims, i.e. to find evidence against them. This is the process commonly carried out by fact-checking organizations, e.g. PolitiFact. In this paper, we introduce a classifier...

Mapping: This narrative may typically include application means like gathering and presenting evidence or providing labels/veracity scores, supported by modeling means like classify/score veracity or evidence retrieval. However, data actors are not specified, and it remains unclear how the debunking outputs will be used.

F.4.6   Assisted Media Consumption

Definition: If the paper proposes to deploy automated fact-checking as an assistive tool for consumers of information, either as a layer adding extra information to social media posts or as a standalone site where claims can be tested.

Example:

Misinformation is a major societal problem, eroding community trust and costing lives by e.g. inducing hesitance to adopt life-saving vaccines. It is therefore of paramount importance that the spread of false information is stopped. Studies have shown that many people adopt beliefs without doing due diligence on the information they receive. Automated fact-checking – that is, the automatic classification of claim veracity – could via e.g., a plugin be deployed to warn social media users about potentially false claims.

Mapping: This narrative may typically feature model owners as social media companies, and data subjects as social media users. The application means can include vague persuasion techniques or aim to limit misinformation by subtly influencing user interaction with content. Alternatively, data actors may be media consumers directly benefiting from these systems.

F.4.7 Scientific Curiosity

Definition: When the authors of the paper justify their projects purely based on scientific curiosity. While differing strongly from the other narratives presented here, this is still a virtue epistemic narrative, concerned with the production of good knowledge.

Example:

Journalistic fact-checking is a difficult task, requiring reasoning about disputed claims that fool sufficiently many humans to warrant professional attention. For systems to mimic fact-checking to a substantial degree, significant semantic understanding is necessary. As such, automated fact-checking is an ideally suited field to develop and test new models for natural language understanding.

Mapping: This narrative may typically involve model owners and data actors who are scientists. The end is to develop knowledge of NLP/language, focusing on advancing scientific understanding rather than practical applications.

F.4.8 Assisted Knowledge Curation

Definition: When the paper proposes fact-checking primarily as a component filtering

the information kept in some curated knowledge vault, including graph-based knowledge bases as well as text-based collections such as Wikipedia.

Example:

Knowledge bases fuel many real-world NLP applications, e.g., question answering. The maintenance of knowledge bases is an expensive process, yet as new facts appear in the world knowledge bases must be kept up-to-date. Automated triple extraction from e.g., news data has been proposed as an alternative to human annotators; yet, the quality remains low. Automated fact-checking systems, which verify facts against trusted knowledge sources, could be used to prevent highly disputed facts from being entered – or ensure that new facts are consistent with the existing knowledge.

Mapping: This narrative may typically include data actors like engineers and curators who use application means to filter system outputs and modeling means such as classify/score veracity, evidence retrieval, or human-in-the-loop techniques. The end is to increase the veracity of published content, ensuring high-quality information in curated knowledge bases.

F.4.9 Adversarial Research

Definition: When the paper proposes the development or use of automated fact-checking (AFC) systems to explore vulnerabilities in misinformation detection, either by generating adversarial examples, repurposing real content in misleading ways, or testing system robustness. The focus is on exposing weaknesses in fact-checking frameworks or demonstrating new, sophisticated threat scenarios (e.g., mismatched image-caption pairs, AI-generated deepfakes).

Mapping: Adversarial research typically occurs when data actors are scientists, who employ modeling means to generate claims and use application means to produce misinformation.

F.4.10 Assisted Internal Fact-Checking

Definition: When the paper proposes to deploy automated fact-checking as an assistive

tool for journalists, deployed internally.

Example:

Research is a fundamental task in journalism, conducted to ensure published information is truthful and to protect the publisher from libel suits. This is a crucial step, which journalists – strained by the advent of the 24-hour news cycle – increasingly skip. Given a trusted source of evidence documents, such as LexisNexis, much of the grunt work of research could be handled by automated fact-checkers, leaving journalists free to tackle the hardest parts, e.g. double-checking information with sources.

Mapping: This narrative may typically feature data actors and data subjects as professional journalists, citizen journalists, or technical writers, using application means that do not supplant human fact-checkers but help to increase the veracity of published content or limit misinformation. This narrative emphasizes internal verification processes within journalistic organizations before content publication.

F.4.11 Automated Content Moderation

Definition: If the paper analyzed proposes a similar content moderation strategy, but instead of assisting human moderators, it suggests replacing them entirely. In this case, the end is to limit the spread of misinformation on social media platforms, the means is to deploy classifiers to truth-tell claims and remove any labeled false, and the actors are the algorithm (as well as, implicitly, the model owners – the executives and engineers at social media companies who deploy and make decisions about such systems).

Example:

Social media is rife with misinformation, eroding community trust and costing lives by e.g. inducing hesitance to adopt life-saving vaccines. One solution is for moderators to remove information deemed false. However, with the number of posts made every day on social networks, this strategy is too costly. In this paper, we develop an automated system for detecting false claims, which can serve as a first line of defense against misinformation.

Mapping: This narrative may typically feature data actors as algorithms, model owners as social media companies, and data subjects as social media users. The application means can also involve automated removal of content to reduce misinformation without human intervention.

F.4.12 Truth-Telling for Law Enforcement

Definition: When the paper proposes fact-checking primarily as a lie detector for use in law enforcement contexts. This could include police work as well as courtroom applications.

Example:

Automatic detection of deception is commonly used in police work via e.g. polygraphs. However, accuracy remains low. We propose that automated fact-checking via NLP could represent a viable alternative.

Mapping: In this narrative, data actors are typically law enforcement, utilizing systems with the end to detect falsehood for law enforcement. These tools support investigations, legal processes, or forensic analysis.

Prediction of the ranking model:
```
{Ranking    explanations    split    by
confidence levels}
```
Using your judgment and reasoning, identify the narrative(s) in the provided excerpt.