Who Plays Which Role?
Protagonist Detection and Classification in Moral Discourse

Mirko Sommer
Department of Computational Linguistics
Heidelberg University
sommer@cl.uni-heidelberg.de

Maria Becker
Department of German Linguistics
Heidelberg University
maria.becker@gs.uni-heidelberg.de

Abstract

Protagonists play a central role in moral discourse by structuring responsibility and authority, yet computational work has largely focused on moral values rather than the actors involved. We address this gap by studying phrase-level protagonist detection and classification in the Moralization Corpus (Becker et al., 2025), a dataset of moral arguments across different text genres. We decompose the task into identifying protagonist mentions and classifying them by what kind of actor they are (e.g., individual or institution) and what function they serve in the moral argument. We compare fine-tuned lightweight models, state-of-the-art NER models, and prompting-based large language models. We further establish human baselines and analyze the impact of contextual information on human and model decisions. Our results show that fine-tuned NER models achieve competitive detection performance at substantially lower cost than prompted large language models, and that role classification benefits more strongly from contextualized prompting. Across tasks, top-performing models reach or exceed human-level performance, highlighting the value of task decomposition for modeling protagonists in moral discourse.
We release our code, predictions, and supplementary material in our project repository.[1]

1 Introduction

Moralizations – arguments that rely on moral values such as *peace* or *freedom* (Becker et al., 2025) – play a central role in public and political communication by articulating normative evaluations, demands, and responsibilities (see Fig. 1 for examples). Beyond identifying moral values or judgments, understanding moralization requires modeling the actors involved: who makes moral demands, who is addressed, and who stands to benefit or suffer. These actors – referred to as protagonists – form a key component of moral discourse and enable fine-grained analyses of moral agency and responsibility in text.

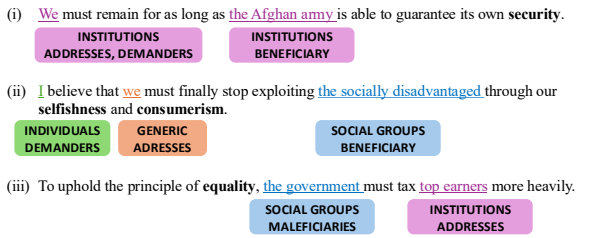


Figure 1: Examples from MORCORP, labeled with protagonist groups and roles (translation from German by the authors). Moral values in **bold**.

Although computational research on moral discourse has predominantly focused on identifying moral values and judgments (see, e.g., Trager et al. (2022); Mirzakhmedova et al. (2024); Falk and Lapesa (2025)), modeling the actors involved is crucial for understanding how morality is constructed and communicated in text. In this paper, we address this gap by systematically modeling protagonists in moralizing discourse at the phrase level, based on the recently released *Moralization Corpus* (MORCORP) (Becker et al., 2025), a large, multi-genre dataset annotated with moral values, demands, and protagonists. We decompose the task into two conceptually distinct stages: (i) detecting textual spans that refer to protagonists, and (ii) classifying these spans according to protagonist group (e.g., INDIVIDUAL, INSTITUTION) and role (e.g., DEMANDER, ADDRESSEE, BENEFICIARY). This decomposition improves transparency, enables targeted evaluation, and allows us to compare fine-tuned lightweight models with state-of-the-art NER models and prompting-based large language models (LLMs) under controlled conditions.

Our study addresses four central **research questions**: **(RQ1)** whether decomposing protagonist

[1] https://github.com/GS-Uni-Heidelberg/Paper-WhoPlaysWhichRole

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 375–392
March 26, 2026 ©2026 Association for Computational Linguistics

modeling into detection and classification subtasks provides advantages over prior all-in-one approaches in terms of performance, transparency, and error localization; **(RQ2)** how effectively protagonist groups and roles can be modeled at the phrase level, and how phrase-level supervision compares to token-level approaches; **(RQ3)** whether fine-tuning smaller models remains a competitive and cost-effective alternative to prompting for specialized linguistic tasks; and **(RQ4)** how different degrees of pre-annotation and contextual information affect classification performance.

In addition to automatic evaluation, we establish multiple human baselines and conduct an annotation experiment to analyze the role of textual context in protagonist annotation. By comparing human and model behavior with and without context, we elucidate systematic differences in how contextual cues are weighted during interpretation.

Overall, our **contributions** are threefold: **(i)** we demonstrate that task decomposition substantially improves end-to-end performance over prior all-in-one approaches, **(ii)** we present a comprehensive comparison of fine-tuning and prompting for protagonist detection and classification, and **(iii)** we provide a detailed human–model analysis of context effects, offering new insights into the strengths and limitations of current language models for discourse-level semantic analysis.

2 Related Work

Modeling Actors in Moral Discourse. Computational research on moral discourse has predominantly focused on identifying moral values and judgments (see eg. Trager et al. (2022); Mirzakhmedova et al. (2024); Landowska et al. (2024); Bulla et al. (2024); Weber-Genzel et al. (2024); Falk and Lapesa (2025); Bulla et al. (2025)) often drawing on Moral Foundations Theory (Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham et al., 2013) to classify moral language in text. However, these approaches typically focus on moral values as isolated labels and offer limited insight into how moral arguments are structured around discourse participants. Only few works address the task of modelling morality in a more holistic, frame based approach (for an overview see Reinig et al. (2024)). Roy et al. (2021, 2022) for example argue that moral labels alone are insufficient to capture meaningful differences in moral reasoning and show that identical moral foundations can be used with different targets to express opposing political positions, underscoring the importance of modeling who is involved in moral discourse, not just which moral values are invoked. Similarly, the Moral Framing In Politics (MFiP) corpus (Rehbein et al., 2025) extends moral framing analysis by annotating narrative roles such as Victim, Villain, Hero, and Beneficiary in German parliamentary debates, though role labeling is not yet treated as a primary modeling objective and left to future work.

The **Moralization Corpus** introduced by Becker et al. (2025) – which builds the basis of our experiments – represents an important step toward more comprehensive moral discourse modeling by conceptualizing moralizations as arguments that invoke moral values to justify demands or positions. Its frame-based annotation scheme captures moral values, demands, and discourse protagonists across diverse German text genres. While this work demonstrates the feasibility of modeling protagonists by prompting LLMs, protagonist detection and classification are treated as auxiliary tasks within a single, unified prompting pipeline, leaving open questions regarding task decomposition, error sources, modeling choices, and context sensitivity. In contrast, we move beyond prior unified modeling approaches by explicitly decomposing protagonist modeling into detection and classification subtasks and by systematically evaluating modeling choices and context effects.

Beyond moral discourse, **modeling actors and their roles** has a long tradition in discourse analysis, for instance, through semantic role labeling (Ruppenhofer et al., 2009; Roth and Lapata, 2015; Bornheim et al., 2024) and participant modeling (Tilk et al., 2016; Ghosh et al., 2023). However, these approaches typically focus on predicate–argument relations and do not capture discourse-level roles specific to moral contexts.

Protagonist detection is also closely related to **Named Entity Recognition** (NER), which aims to identify and classify entity mentions such as persons, organizations, and locations. While NER is a mature area of NLP with extensive surveys and benchmarks (e.g. Zhang et al. (2025); Seow et al. (2025) for the latest ones; see also Tong et al. (2025) for NER with LLMs and prompting techniques), classical NER focuses on ontological entity types and does not account for discourse roles or pragmatic functions, which are central to our approach.

3 Dataset

Moralization Corpus (MORCORP). For our analyses, we use MORCORP (Becker et al., 2025), a multi-genre dataset in German designed to analyze moralizations in discourse based on a frame-based annotation scheme that captures the constitutive elements of moralizations – moral values, demands, and discourse protagonists.

The dataset comprises 11,503 short paragraphs of one to five sentences from several text genres, such as political debates, news articles, and online discussions. It includes both moralizing and negative instances – paragraphs that refer to moral values without any persuasive intent (a key characteristic of moralizations; see Becker et al. (2025)). For our approaches, we use only the moralizing instances (18% of the paragraphs), adhering to the train/test/dev split of Becker et al. (2025) (70:15:15 train/test/dev, with balanced genre distribution).

Protagonist Annotations. In the dataset, all protagonists (individuals or groups) have been annotated on the *phrase level* in a multi-step annotation procedure (for details, see Becker et al. (2025) on the phrase level[2] along two dimensions: group type and role (see Fig. 1 for annotated examples from MORCORP.). **Group types** include the sublabels INDIVIDUALS (e.g. *Angela Merkel*), GENERIC (references to humans, such as *the people, citizens*), INSTITUTIONS/ORGANIZATIONS (e.g. *the democrats, the stakeholders*), and SOCIAL GROUPS (e.g. *parents, homeless people*). **Roles** capture the role of the respective protagonist within the moralization and distinguish between the person who is moralizing (DEMANDER), the person who is the target of the demand (ADRESSEE), and the person who would benefit (BENEFICIARY) or be disadvantaged (MALEFICIARY) from the demand. While group classification is a single-label task, role classification allows multiple labels for the same entity, as a protagonist may simultaneously occupy several relational positions (e.g., an inclusive 'we' can function as both DEMANDER and ADDRESSEE).

Data Statistics. The most frequent protagonist roles in the dataset are beneficiaries (0.65 per instance) and addressees (0.64), followed by demanders (0.42), while maleficiaries are rare (0.10). In terms of group distribution, institutions (32% of all annotated protagonists) and social groups (30%) dominate, followed by individuals (20%) and generic human references (15%).

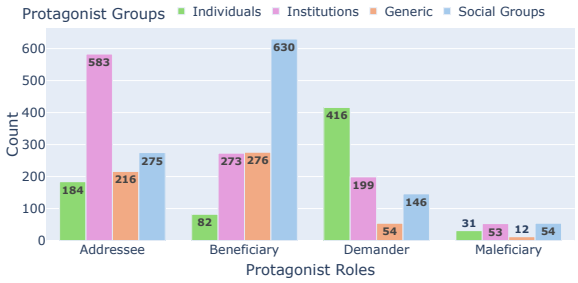


Figure 2: Co-occurrences of protagonist groups and roles in MORCORP. Absolute counts.

Co-occurrence patterns of groups and roles (Fig. 2) reveal systematic structure: demanders are typically individuals, addressees are predominantly institutions, and beneficiaries are mainly social groups or generic humans. Maleficiaries, though infrequent, are primarily institutions or social groups, suggesting that moral demands often target or disadvantage collective actors.

Prior Experiments. Becker et al. (2025) evaluate several LLMs on tasks derived from MORCORP, including moralization detection, value classification, and the detection and classification of protagonists. While their primary focus lies on moralization detection across diverse genres, the remaining tasks are treated mainly as probing or feasibility experiments rather than as fully developed analytical components. All tasks are addressed within a single chain-of-thought (CoT) prompting setup, in which moralization detection precedes and subsumes the identification of moral values and protagonists.

Protagonist detection and classification are evaluated using a SemEval-2013 NER-style strict and partial matching setup (Segura-Bedmar et al., 2013). Performance is low: even the best configuration (gpt-5-mini with few-shot prompting) achieves F1 scores of only 0.19 for groups and 0.18 for roles, highlighting the task's complexity. We build on these feasibility studies by systematically modeling protagonists using both prompting and fine-tuning approaches. Crucially, we decompose the task into protagonist detection (§ 4) and protagonist classification (§ 5), improving transparency and interpretability. This separation allows errors to be more clearly localized and reduces overall task complexity by disentangling token-level detection from higher-level discourse interpretation.

[2]Phrase-level protagonist annotation differs from standard NER, which relies on well-defined entity boundaries. In MORCORP, span boundaries are phrase-based, leading to increased variability in how spans are delimited.

4 Detecting Protagonists

We first focus on predicting textual spans that refer to protagonists. Specifically, the goal is to identify and extract contiguous text spans denoting actors or groups involved in the moral discourse, irrespective of their subsequent classification by group affiliation or role.

4.1 Models

To explore trade-offs between supervision, computational cost, and modeling flexibility, we examine three approaches to protagonist detection: fine-tuning general-purpose language models, adapting task-specific NER models, and prompting LLMs. This comparison contrasts lightweight supervised methods with zero- and few-shot prompting that require no task-specific training data.

Fine-tuned Base Language Models. We fine-tune *bert-base-german-cased*, a general-purpose German language model based on BERT (Devlin et al., 2019), using a sequence labeling approach. A key advantage is its low computational cost, with fine-tuning and evaluation requiring only modest hardware (≈8 hours on one NVIDIA 1080 GPU).

Fine-tuned NER Models. As discussed in § 2, the task of protagonist detection is related to named entity recognition (NER). We therefore experiment with pre-trained German NER models, specifically *ner-german-large* (Schweter and Akbik, 2020) and *ner-german* (Akbik et al., 2018). We finetune the models on the protagonist annotations in MORCORP using a standard BIO tagging scheme, labeling protagonist spans as B-*TYPE*/I-*TYPE* and all other tokens as O.

Since off-the-shelf NER models assume a fixed inventory of entity labels (e.g., PER, ORG, LOC), we map each protagonist class in MORCORP to the closest corresponding NER label (e.g., INSTITUTION → ORG). This mapping ensures compatibility with the models' output space and avoids inconsistent gradients during fine-tuning. The complete mapping is provided in A.5.

Like the BERT models, these NER models require only modest resources, with fine-tuning taking ≈10 hours on a single NVIDIA 1080 GPU.

Base NER Models. We compare the performance of the fine-tuned NER models with their untuned base counterparts to assess how much task-specific fine-tuning improves the transfer of NER representations to the protagonist detection task.

Prompting Approaches. We prompt LLMs for protagonist detection using different prompting strategies: (i) ***pd_basic_0shot***, which provides a minimal instruction; (ii) ***pd_cot_0shot***, which incorporates step-by-step reasoning, along with a detailed description of what constitutes a protagonist in moral contexts, derived from the annotation guidelines provided by Becker et al. (2025); (iii) ***pd_cot_10shot***, which extends *pd_cot_0shot:* by adding 10 example sentences with annotated protagonist spans; and (iv) ***pd_cot_10shot_def***, the most comprehensive setting, which combines *pd_cot_10shot* with a detailled definition of moralizations and invloved actors.

We conduct experiments with two instruction-tuned LLMs that differ in architecture and context window capacity: GPT-5-mini-2025-08-07 and Claude-3.5-Haiku-20241022.

4.2 Human Baseline

Finally, we establish a human agreement baseline for interpreting model performance. Since the protagonist annotations in MORCORP were created within a multi-stage annotation process, where protagonist spans were refined iteratively rather than annotated in parallel, inter-annotator agreement scores are not available for protagonists. To approximate human-level performance, a trained linguistics student independently annotates all protagonist spans, along with their group and role labels, on a subset of 50 paragraphs from MORCORP, strictly following the original guidelines from Becker et al. (2025). We then evaluate the resulting annotations against the MORCORP annotations of protagonists (referred to as *gold labels*) using the same metrics as for the automatic detection models (see §4.3).

4.3 Metrics and Results

Metrics. We evaluate our models using micro-averaged F1 scores under strict and partial matching (for P/R scores see A.6). Under strict matching, predictions must exactly match the gold spans, while partial matching assigns reduced credit to overlapping spans (see A.4 for formal definitions). We assess pairwise statistical significance using the 5×2 cross-validation test (Dietterich, 1998); full significance matrices are reported in A.8.

Results are displayed in Table 1. Overall, they reveal clear performance and efficiency trade-offs across modeling approaches: Fine-tuned NER models – particularly flair-ner-german-large – achieve performance comparable to prompting with GPT-5-mini while requiring substantially fewer compu-

	model name	experiment	F1 (micro-avg) strict	F1 (micro-avg) partial
base LMs	bert-base-german-cased	fine_tuned	0.4325	**0.5288**
ner	flair-ner-german	base	0.1713	0.2521
		fine_tuned	0.4137	0.4413
	flair-ner-german-large	base	0.1695	0.2548
		fine_tuned	**0.4783**	**0.5375**
prompting	claude-3-5-haiku	pd_basic_0shot	0.3350	0.4667
		pd_cot_0shot	0.3586	0.5134
		pd_cot_10shot	0.3896	0.4955
		pd_cot_10shot_def	0.3896	0.5090
	gpt-5-mini	pd_basic_0shot	0.3504	0.3822
		pd_cot_0shot	**0.4789**	0.5163
		pd_cot_10shot	**0.4865**	0.5200
		pd_cot_10shot_def	**0.4882**	0.5211
human	human	annotation_context	0.4405	0.4835

Table 1: Results of protagonist *detection*. Best-performing models and scores not significantly different ($p \geq 0.05$) from the top model in **bold**.

tational resources, highlighting the effectiveness of lightweight supervised approaches. Across settings, our strongest models reach or exceed the human baseline, indicating that protagonist detection can be performed at near-human reliability. Partial span matching yields notable gains only for untuned base and NER models, suggesting that phrase boundary detection is not a primary source of error for fine-tuned or prompted models. Moreover, enriching prompts with examples or detailed definitions does not yield significant improvements, indicating diminishing returns from increasing prompt complexity. Finally, fine-tuning consistently yields significant gains for NER models, and model size and capability matter: GPT-5-mini outperforms Claude-3.5-Haiku, and flair-ner-german-large surpasses its smaller counterpart.

5 Classifying Protagonists

Next, we address the task of *classifying* protagonist phrases by group and role attributes. According to MORCORP (see § 3), we define group classification as a single-label task, whereas role classification is multi-label since a protagonist may fulfill multiple roles.

Beyond model comparisons, we contrast *context-free* and *context-aware* protagonist classification. In the context-free setting, models predict group or role labels from the protagonist phrase alone; in the context-aware setting, they additionally receive the full paragraph containing the marked span. This design isolates the contribution of contextual information. We hypothesize that group classification is largely context-independent, as group membership is typically lexically encoded (e.g., parents → SOCIAL GROUP), whereas role classification is inherently context-dependent, since roles such as demander or beneficiary are defined by discourse relations.

5.1 Setup 1: Oracle

In this setup, models receive gold protagonist spans from MORCORP, constituting an oracle setting. This decouples classification from detection, isolates labeling performance, and establishes an upper bound on classification independent of span detection errors.

5.1.1 Models

Statistical Baselines. We use a **rule-based n-gram classifier** as a baseline, which associates n-grams with their most frequent group or role labels and predicts via vote aggregation. Relying only on shallow lexical cues from protagonist phrases, without context, serves as a lower bound for context-free classification. As a stronger baseline, we train a **random forest** on TF–IDF-weighted n-gram features from protagonist phrases using scikit-learn[3]. Without incorporating context, this remains a surface-level lexical baseline but captures richer cues than the n-gram model. For multi-label role classification, we use a one-vs-rest setup with a separate binary classifier for each role.

Fine-tuned Base Language Models. Analogous to protagonist detection (§ 4), we fine-tune the BERT-base model for group and role classification in two configurations. In **fine_tuned**, the model acts as a sequence classifier, treating each protagonist phrase independently and relying only on phrase-internal lexical and morphosyntactic features. In **fine_tuned + masked_context**, the model is trained as a token classifier over the full paragraphs, with loss computed only on protagonist spans and span-level predictions obtained via mean pooling. This setup incorporates contextual cues beyond the surface form, while remaining feasible on modest hardware.

Prompting Approaches. As in protagonist detection, we evaluate several prompting strategies for group and role classification: (i) ***pd_basic_0shot*** with a minimal instruction; (ii) ***pd_cot_0shot***, adding step-by-step reasoning and detailed label descriptions; (iii) ***pd_cot_10shot***, which includes 10 annotated example phrases; (iv) ***pd_cot_10shot_context***, additionally providing the full paragraph for each protagonist phrase; and (v) ***pd_cot_10shot_context+def***, which further

[3] https://scikit-learn.org

	model name	experiment	f1	confidence
statistical baselines	ngram rule-based	fine_tuned	0.4473	-
	random forest	fine_tuned	0.5158	0.6161
base LMs	bert-base-german-cased	fine_tuned	**0.7197**	0.9173
		fine_tuned + masked_**context**	**0.7162**	0.9015
prompting	claude-3-5-haiku	pgc_basic_0shot	0.6037	0.9018
		pgc_cot_0shot	0.6757	0.8948
		pgc_cot_10shot	0.6766	0.8733
		pgc_cot_10shot_**context**	0.6819	0.8796
		pgc_cot_10shot_**context**+def	0.6819	0.8701
	gpt-5-mini	pgc_basic_0shot	0.6564	0.9025
		pgc_cot_0shot	0.7030	0.9018
		pgc_cot_10shot	0.7047	0.8935
		pgc_cot_10shot_**context**	**0.7452**	**0.9405**
		pgc_cot_10shot_**context**+def	**0.7408**	**0.9412**
human	human 1	annotation	0.6569	0.6759
		annotation_**context**	0.5839	0.8606
	human 2	annotation	0.5620	0.8044
		annotation_**context**	0.7080	0.9234

(a) Protagonist *group* classification.

	model name	experiment	f1	confidence
statistical baselines	ngram rule-based	fine_tuned	0.4368	-
	random forest	fine_tuned	0.4747	0.6160
base LMs	bert-base-german-cased	fine_tuned	0.5588	0.6697
		fine_tuned + masked_**context**	0.5895	0.7823
prompting	claude-3-5-haiku	prc_basic_0shot	0.4089	0.6666
		prc_cot_0shot	0.4443	0.6921
		prc_cot_10shot	0.2476	0.6554
		prc_cot_10shot_**context**	0.2851	0.8167
		prc_cot_10shot_**context**+def	0.2846	0.8030
	gpt-5-mini	prc_basic_0shot	0.3585	0.6735
		prc_cot_0shot	0.3432	0.4882
		prc_cot_10shot	0.3601	0.5074
		prc_cot_10shot_**context**	0.6204	**0.8811**
		prc_cot_10shot_**context**+def	**0.6426**	**0.8816**
human	human 1	annotation	0.5177	0.2978
		annotation_**context**	0.6494	0.7853
	human 2	annotation	0.4752	0.5620
		annotation_**context**	0.7203	0.8511

(b) Protagonist *role* classification.

Table 2: Results for protagonist *classification*. Best-performing models and scores not significantly different ($p \geq 0.05$) from the top model are shown in **bold**. All values are micro-averaged.

supplies detailled definitions of moralization and involved actors. We use the same LLMs as in § 4.1 (GPT-5-mini and Claude-3.5). All prompts are available in our repository.

5.1.2 Human Baseline

To obtain an upper human baseline for the oracle setting, we conduct an additional annotation experiment on the same 50-paragraph MORCORP subset used in § 4.2. Two trained linguistics annotators independently assign group and role labels to all protagonist phrases in two subsequent settings: First, they receive only isolated phrases (context-free setting), while in the context-aware setting, annotators additionally receive the surrounding context, matching the input conditions of contextualized models. In addition, annotators provide confidence ratings from 1-10 in both conditions to assess the effect of context on certainty.

We evaluate annotations against one another and against other models, using PABAK (Byrt et al., 1993) as the agreement metric, and against the gold labels in MORCORP using the same metrics as for automatic models, reporting the resulting human–gold agreement as a baseline.

5.1.3 Metrics and Results

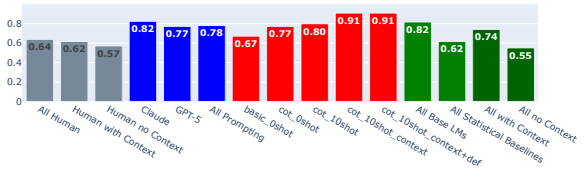
Metrics. Evaluation is again performed using micro-averaged F1-scores (for P/R and class-specific scores see A.7 and A.6). Statistical significance testing follows the procedure in § 4.3 (see A.8 for significance matrices).

In addition, we assess model (un)certainty via confidence scores, reported as the average confidence of all predictions. For BERT models, NER models, and random forests, confidence corresponds to the predicted class probabilities. Prompting-based models are explicitly instructed to report confidence, following prior work suggesting that language models can estimate their own correctness when prompted (Kadavath et al., 2022).[4]

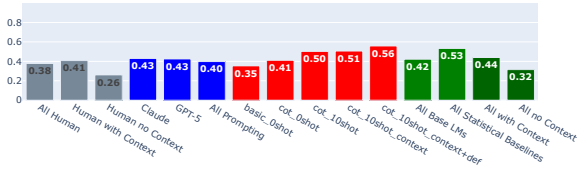
Results. Tables 2a and 2b show that our strongest models match or exceed the human baseline for both group and role classification and substantially outperform statistical baselines. For group classification, fine-tuned models achieve performance comparable to prompting at much lower cost, highlighting the effectiveness of lightweight supervised models. In contrast, role classification proves more challenging and benefits more strongly from prompting, with some prompting configurations outperforming our fine-tuned models. As hypothesized, context yields larger gains for roles than for groups, reflecting its stronger dependence on discourse-level cues. Notably, GPT benefits from added context and moralization definitions, whereas Claude's performance declines, indicating model-specific sensitivities to prompt complexity.

Confidence Scores. Confidence patterns in Tables 2b and 2a broadly mirror performance. Stronger models (esp. GPT) show higher confidence, whereas weaker baselines exhibit lower and more variable scores. Context consistently increases confidence, with a stronger effect for role than for group classification, reflecting the greater discourse dependence of roles, which are defined by relations and actions in context rather than phrase-internal properties. Group classification yields higher and more stable confidence overall, a pattern also observed for human annotators.

[4]Such self-reported confidence should be interpreted cautiously, as it reflects subjective estimates rather than calibrated probabilities and is sensitive to prompt and model design; we therefore treat it as a relative indicator of uncertainty.



(a) Mean PABAK scores for *group* classification.



(b) Mean PABAK scores for *role* classification.

Figure 3: IAA scores for protagonist classification.

IAA Scores. Fig. 3a and 3b report PABAK agreement for protagonist group and role classification, comparing humans and models across system families and context conditions. This comparison assesses proximity to human performance, variation across modeling paradigms, and the impact of contextual information.

For group classification, agreement is consistently high across humans and top-performing models, with minimal differences between context-free and context-aware settings, indicating that group membership is largely encoded in the protagonist phrase itself. In contrast, role classification shows lower agreement and greater variability. Both humans and models benefit from context, reflecting the discourse-dependent and ambiguous nature of role assignment. The greater spread across model types highlights the task's conceptual complexity. Notably, even human agreement remains moderate, underscoring that role classification is subject to interpretation.

5.2 Setup 2: Pipeline

To reflect realistic use, we evaluate our models in a pipeline setting where protagonist spans are not given a priori. Protagonist detection and subsequent group and role classification are performed sequentially, using the output of our best detection models as input to the best classification models (see A.3 for our selection process). We compare a best prompting pipeline and a best fine-tuned pipeline against the strongest baseline from Becker et al. (2025), enabling assessment of end-to-end performance and cost–performance trade-offs. The selected model and experiment names are available in Table 9 (A.3).

	PGC pipeline f1	PRC pipeline f1
baseline (Becker et al., 2025)	0.1868	0.1800
best fine-tuning	0.3543	0.2865
best prompting	**0.5628**	**0.4846**
human	0.3043	0.1836

Table 3: Results of the pipeline approach, where protagonist detection is followed by protagonist group classification (PGC) or protagonist role classification (PRC). Reported scores for the strict evaluation variant, micro-averaged. Best-performing model in **bold**.

5.2.1 Human Baseline

To establish a human baseline for the pipeline setting, we reuse the protagonist annotations from the 50-paragraph MORCORP subset (§ 4.2). As before, we evaluate these annotations against the gold labels using the same metrics as for automatic models (see below) and report the resulting human–gold agreement as the baseline.

5.2.2 Metrics and Results

Metrics. We evaluate our pipeline following the setup of Becker et al. (2025) using SemEval-2013 NER-style strict and partial matching (Segura-Bedmar et al., 2013) (see A.4 for details and definitions). This enables a direct comparison between the all-in-one (or single-stage) approach of Becker et al. (2025) and our decomposed detection–classification pipeline.

Results. As shown in Table 3, all of our pipeline configurations substantially outperform the all-in-one results of Becker et al. (2025), demonstrating the effectiveness of decomposing protagonist detection and classification into separate stages. Notably, our strongest pipelines exceed the human baseline for both group and role classification, indicating robust end-to-end performance under realistic conditions. While the best prompting pipeline achieves higher scores than the best fine-tuned pipeline, fine-tuned models remain attractive due to their substantially lower computational and monetary costs.

5.3 Analysis of Context Effects on Labels

To analyze the effect of contextual information, we compare instances in which humans or models (focusing on prompting approaches) change their label assignments once context is provided with those in which labels remain stable.[5] We hypothesize that instances without label changes indicate

[5]Please note that we are focusing on mechanisms of label change rather than error analysis.

the feasibility of context-free labeling. We enrich the 50-paragraph MORCORP subset annotated for protagonist groups and roles in a context-free and in a subsequent context-aware setting (see§5.1.2): For each protagonist phrase, we manually assign linguistic features that might have a decisive effect on model and human labelling decisions. The features are derived from bottom-up data inspection and linguistically motivated hypotheses (e.g., that longer phrases or generic references should be classified reliably without context, as opposed to, e.g., pronoun phrases). We include **semantic features**: generic reference, named entities, political reference, and sentiment; and **syntactic features**: phrase type and token length (see A.10 for details).

Table 15 (A.10) compares human annotations with Claude and GPT predictions for group and role classification. **Overall**, across humans and models, label changes are more frequent for role than for group classification, confirming the stronger context dependence of roles. Humans revise labels substantially more often than models (30.7% vs. 13.1–10.2% for groups; 52.6% vs. 18.2% for roles), while GPT exhibits an exceptional pattern, changing role labels in 69.3% of cases, indicating a strong reliance on context for role classification.

For **group classification**, human label changes are primarily associated with syntactic properties: noun phrases are overrepresented among changed labels (71.6% vs. 53.7%), whereas pronouns are more stable (14.2% vs. 35.3%). In contrast, models rely more on semantic cues; for example, in Claude's predictions, phrases with negative sentiment are much more frequent among stable than changed labels (15.1% vs. 1.6%).

For **role classification**, human label changes are mostly driven by semantic cues such as political references or negative sentiment, reflecting the inherently discourse-dependent nature of roles. In contrast, the models exhibit less systematic behavior: for example, Claude frequently changes role labels for generic expressions (28% of changed cases), a pattern that appears linguistically implausible and suggests weaker alignment with human role-assignment strategies.

Overall, while LLMs broadly mirror human sensitivities to context, they differ systematically in how they weight contextual cues. Humans revise labels selectively and in linguistically interpretable ways, whereas models exhibit inconsistent behavior and, at times, lack semantic plausibility.

6 Discussion and Conclusion

In this paper, we presented a systematic study of phrase-level protagonist detection and classification in moral discourse. By decomposing protagonist modeling into detection and classification, we disentangle sources of error, improve interpretability, and outperform prior feasibility-oriented end-to-end approaches on the Moralization Corpus.

Our results reveal clear trade-offs between fine-tuning and prompting. For protagonist detection and group classification, fine-tuned lightweight models, particularly adapted NER models, achieve performance comparable to prompting-based LLMs at a fraction of the computational and monetary cost. In contrast, role classification is more challenging and benefits more strongly from contextualized prompting, reflecting its discourse-dependent nature. Across tasks, our strongest models reach or exceed human-level performance, indicating that protagonist modeling is feasible at near-human reliability.

Beyond performance, our analysis reveals systematic differences in how humans and LLMs exploit context. Human annotators revise labels selectively and in linguistically interpretable ways, guided by syntactic and semantic cues, whereas LLMs show mixed patterns with limited semantic plausibility. This suggests that, despite strong overall performance, LLMs may rely on surface-level or overly broad contextual cues rather than fully internalizing task-specific distinctions. This indicates that higher performance does not necessarily imply human-like reasoning, underscoring the need for complementary human-centered analyses.

Overall, our study shows that task decomposition and targeted modeling choices are crucial for advancing computational analyses of moral discourse. Our findings underscore the importance of modeling discourse actors for understanding how moral arguments are constructed and communicated. Future work should explore multilingual extensions, improved calibration of contextual reasoning, and tighter integration of discourse structure.

Although our analysis centers on moral discourse, discourse actors and role structures extend beyond this domain. Comparable role structures arise in other domains, such as policy narratives (Schläufer et al., 2022) with hero, victim, and villain roles, making our framework a promising foundation for modeling actors across argumentative and narrative discourse.

Limitations

Our study has several limitations. First, all experiments are conducted on German data from the Moralization Corpus, which limits the generalizability of our findings to other languages and cultural contexts. Moral discourse and the realization of protagonist roles may differ substantially across languages and discourse traditions; therefore, future work should evaluate multilingual and cross-cultural extensions.

Second, while we compare fine-tuned lightweight models and prompting-based large language models, we restrict fine-tuning to relatively small and mid-sized models for reasons of computational efficiency. Fine-tuning larger transformer models may further improve performance, particularly for role classification, but was beyond the scope of this work.

Third, our pipeline evaluation selects only a subset of promising detection–classification combinations based on approximated performance. A full exploration of all possible pipeline configurations could yield additional insights into error propagation and robustness.

Finally, although we establish multiple human baselines and conduct a dedicated annotation study, protagonist role classification remains inherently ambiguous, as reflected in moderate inter-annotator agreement. Some disagreement, therefore, reflects genuine interpretive variability rather than model error, which should be taken into account when interpreting absolute performance scores.

Ethical considerations

This work analyzes moralizing discourse, which frequently occurs in politically and socially sensitive contexts. While our models do not generate new content, they may nonetheless reproduce biases present in the underlying data, such as overrepresenting certain social groups as addressees or maleficiaries in moral arguments. We therefore caution against deploying such models for normative judgments or automated decision-making without careful human oversight.

All experiments are conducted on an existing, publicly available dataset collected and annotated in prior work (Becker et al., 2025). No personally identifiable information is introduced beyond what is already present in the source texts, and we do not annotate or infer private attributes of individuals.

Trained annotators performed human annotation following detailed guidelines and with explicit awareness of the task's interpretive nature. Annotators were not exposed to model outputs, minimizing confirmation bias. Nevertheless, subjective judgments are unavoidable in role annotation, and our analysis explicitly treats disagreement as an object of study rather than as an annotation error.

Finally, we report computational costs and emphasize lightweight fine-tuned models as viable alternatives to large-scale prompting. This contributes to more sustainable and accessible NLP research by reducing reliance on resource-intensive systems.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Maria Becker, Mirko Sommer, Lars Tapken, Yi Wan Teh, and Bruno Brocai. 2025. The moralization corpus: Frame-based annotation and analysis of moralizing speech acts across diverse text genres. *Preprint*, arXiv:2512.15248.

Tobias Bornheim, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski. 2024. Speaker attribution in German parliamentary debates with QLoRA-adapted large language models. *Journal for Language Technology and Computational Linguistics*, 37(1):1–13.

Luana Bulla, Stefano De Giorgis, Misael Mongiovì, and Aldo Gangemi. 2025. Large language models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609.

Luana Bulla, Aldo Gangemi, and Misael Mongiovì. 2024. Do language models understand morality? towards a robust detection of moral content. In *Value Engineering in Artificial Intelligence*, pages 98–113, Cham. Springer Nature Switzerland.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921.

Sayontan Ghosh, Mahnaz Koupaee, Isabella Chen, Francis Ferraro, Nathanael Chambers, and Niranjan Balasubramanian. 2023. PASTA: A dataset for modeling PArticipant STAtes in narratives. *Transactions of the Association for Computational Linguistics*, 11:1283–1300.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jonathan Haidt and Jesse Graham. 2007. When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Alina Landowska, Katarzyna Budzynska, and He Zhang. 2024. Quantitative and qualitative analysis of moral foundations in argumentation - argumentation.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. The touché23-ValueEval dataset for identifying human values behind arguments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.

Ines Rehbein, Ines Reinig, and Simone Paolo Ponzetto. 2025. Moral framing in politics (MFiP): A new resource and models for moral framing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34631–34651, Suzhou, China. Association for Computational Linguistics.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A survey on modelling morality for text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand. Association for Computational Linguistics.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards few-shot identification of morality frames using in-context learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196, Abu Dhabi, UAE. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. SemEval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 106–111, Boulder, Colorado. Association for Computational Linguistics.

Christopher Schläufer, Jan Künzler, Michael D. Jones, and 1 others. 2022. The narrative policy framework: A traveler's guide to policy stories. *Politische Vierteljahresschrift*, 63:249–273.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition. *Preprint*, arXiv:2011.06993.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Wei Li Seow, Iti Chaturvedi, Andrew Hogarth, and 1 others. 2025. A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58:315.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182, Austin, Texas. Association for Computational Linguistics.

Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. EvoPrompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5136–5153, Abu Dhabi, UAE. Association for Computational Linguistics.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazazian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Lopez Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Eva Luis Álvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus. *ArXiv*, abs/2208.05545.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025. A survey of generative information extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

A Appendix

A.1 Label Translations

The Moralization Corpus uses German labels for the protagonist groups and roles. We translated those labels into English for this paper. In our prompts, code, and repository, we use the German labels, the translations are shown in Table 4 for group classification and in Table 5 for role classification.

Dataset (German)	Paper (English)
Individuum	Individual
Institution	Institution
Menschen	Generic
Soziale Gruppe	Social Group
Other	Other

Table 4: Translation of German dataset labels into English labels used in this paper for group classification.

Dataset (German)	Paper (English)
Adressat:in	Addressee
Benefizient:in	Beneficiary
Forderer:in	Demander
Malefizient:in	Maleficiary
Bezug unklar	Unclear

Table 5: Translation of German dataset labels into English labels used in this paper for role classification.

A.2 Language model identifiers

Table 6 shows the short names used in this paper and their corresponding full model version identifiers.

Short Name	Model Version
gpt-5-mini	gpt-5-mini-2025-08-07
claude-3-5-haiku	claude-3-5-haiku-20241022

Table 6: Language models and corresponding version identifiers.

A.3 Selection of best models for our pipeline

To identify the best-performing models for inclusion in our pipeline, we use the results from Section 4, Section 5.1, and the following formula:

$$\text{Pipeline Rec} = \text{Recall}_{\text{PD}} \cdot \text{Recall}_{\text{PC}}$$

$$\text{Pipeline Prec} = \text{Precision}_{\text{PD}} \cdot \text{Precision}_{\text{PC}}$$

$$\text{Pipeline F1} = \frac{2 \cdot \text{Pipeline Prec} \cdot \text{Pipeline Rec}}{\text{Pipeline Pre} + \text{Pipeline Rec}}$$

These approximations assume that classifier performance on predicted spans and on gold spans is independent, and that detection and classification errors are also independent. Both assumptions are imperfect. Future work should therefore investigate all pipeline configurations, not only those that yield the best approximated results. Table 7 and 8 show the top-20 approximation results for both pipelines.

Table 9 shows our selected models for our pipeline experiments. (see 3).

A.4 Formal Definition of Pipeline and Protagonist Detection Metrics

For evaluating the pipeline and the protagonist detection (Section 4.3 and 5.2.2), we adapted the evaluation framework for NER models as defined in SemEval 2013 - 9.1 task (Segura-Bedmar et al., 2013) and applied by Becker et al. (2025).

For protagonist detection, we reimplemented the evaluation procedure from scratch and reran all experiments. We adaptoed a span-agnostic approach, where predicted and gold protagonist phrases are compared without considering their position, frequency, or category. By disregarding these aspects, we simplified the matching process while

Rank	Detection Model (Experiment)	Group Model (Experiment)	Pipeline F1
1	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (pgc_cot_10shot_context)	**0.3638**
2	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (pgc_cot_10shot_context)	0.3625
3	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (pgc_cot_10shot_context+def)	0.3617
4	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (pgc_cot_10shot_context+def)	0.3604
5	gpt-5-mini (pd_cot_0shot)	gpt-5-mini (pgc_cot_10shot_context)	0.3568
6	flair-ner-german-large (fine_tuned)	gpt-5-mini (pgc_cot_10shot_context)	0.3564
7	gpt-5-mini (pd_cot_0shot)	gpt-5-mini (pgc_cot_10shot_context+def)	0.3547
8	flair-ner-german-large (fine_tuned)	gpt-5-mini (pgc_cot_10shot_context+def)	0.3543
9	gpt-5-mini (pd_cot_10shot_def)	bert-base-german-cased (fine_tuned)	0.3514
10	gpt-5-mini (pd_cot_10shot)	bert-base-german-cased (fine_tuned)	0.3501
11	gpt-5-mini (pd_cot_10shot_def)	bert-base-german-cased_with_mask (fine_tuned)	0.3497
12	gpt-5-mini (pd_cot_10shot)	bert-base-german-cased_with_mask (fine_tuned)	0.3484
13	gpt-5-mini (pd_cot_0shot)	bert-base-german-cased (fine_tuned)	0.3446
14	flair-ner-german-large (fine_tuned)	bert-base-german-cased (fine_tuned)	0.3442
15	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (pgc_cot_10shot)	0.3441
16	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (pgc_cot_0shot)	0.3432
17	gpt-5-mini (pd_cot_0shot)	bert-base-german-cased_with_mask (fine_tuned)	0.3430
18	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (pgc_cot_10shot)	0.3429
19	flair-ner-german-large (fine_tuned)	bert-base-german-cased_with_mask (fine_tuned)	0.3425
20	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (pgc_cot_0shot)	0.3420

Table 7: Top 20 approximated pipeline configuration results, where protagonist detection (PD) is followed by protagonist group classification (PGC).

Rank	Detection Model (Experiment)	Role Model (Experiment)	Pipeline F1
1	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (prc_cot_10shot_context+def)	**0.3097**
2	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (prc_cot_10shot_context+def)	0.3084
3	flair-ner-german-large (fine_tuned)	gpt-5-mini (prc_cot_10shot_context+def)	0.3053
4	gpt-5-mini (pd_cot_0shot)	gpt-5-mini (prc_cot_10shot_context+def)	0.3035
5	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (prc_cot_10shot_context)	0.2990
6	gpt-5-mini (pd_cot_10shot)	gpt-5-mini (prc_cot_10shot_context)	0.2978
7	flair-ner-german-large (fine_tuned)	gpt-5-mini (prc_cot_10shot_context)	0.2948
8	gpt-5-mini (pd_cot_0shot)	gpt-5-mini (prc_cot_10shot_context)	0.2931
9	gpt-5-mini (pd_cot_10shot_def)	bert-base-german-cased_with_mask (fine_tuned)	0.2839
10	gpt-5-mini (pd_cot_10shot)	bert-base-german-cased_with_mask (fine_tuned)	0.2827
11	bert-base-multilingual-cased (fine_tuned)	gpt-5-mini (prc_cot_10shot_context+def)	0.2823
12	flair-ner-german-large (fine_tuned)	bert-base-german-cased_with_mask (fine_tuned)	0.2800
13	gpt-5-mini (pd_cot_0shot)	bert-base-german-cased_with_mask (fine_tuned)	0.2782
14	bert-base-german-cased (fine_tuned)	gpt-5-mini (prc_cot_10shot_context+def)	0.2762
15	bert-base-multilingual-cased (fine_tuned)	gpt-5-mini (prc_cot_10shot_context)	0.2726
16	flair-ner-german (fine_tuned)	gpt-5-mini (prc_cot_10shot_context+def)	0.2672
17	gpt-5-mini (pd_cot_10shot_def)	bert-base-german-cased (fine_tuned)	0.2669
18	bert-base-german-cased (fine_tuned)	gpt-5-mini (prc_cot_10shot_context)	0.2667
19	gpt-5-mini (pd_cot_10shot)	bert-base-german-cased (fine_tuned)	0.2657
20	flair-ner-german-large (fine_tuned)	bert-base-german-cased (fine_tuned)	0.2643

Table 8: Top 20 approximated pipeline configuration results, where protagonist detection (PD) is followed by protagonist role classification (PRC).

	PD model (experiment)	**PGC model** (experiment)
baseline (Becker et al., 2025)	gpt-5-mini (cot_json_10shot)	gpt-5-mini (cot_json_10shot)
best fine-tuning	flair-ner-german-large (fine_tuned)	bert-base-german-cased (fine_tuned)
best prompting	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (pgc_cot_10shot_context)

(a) Protagonist group classification (PGC) pipeline model and experiment configurations.

	PD model (experiment)	**PRC model** (experiment)
baseline (Becker et al., 2025)	gpt-5-mini (cot_json_10shot)	gpt-5-mini (cot_json_10shot)
best fine-tuning	flair-ner-german-large (fine_tuned)	bert-base-german-cased (fine_tuned + masked context)
best prompting	gpt-5-mini (pd_cot_10shot_def)	gpt-5-mini (prc_cot_10shot_context+def)

(b) Protagonist role classification (PRC) pipeline model and experiment configurations.

Table 9: Model and experiment configurations for the pipeline approaches, where protagonist detection (PD) is followed by protagonist group classification (PGC) or protagonist role classification (PRC).

maintaining the exact formulas for precision, recall, and F1 score. Consequently, the reported performance scores differ from the strict and partial binary-protagonist results reported in Becker et al. (2025).

In contrast, for the pipeline approach, we reused the original evaluation code of Becker et al. (2025), which uses the Python library nervaluate[6] to compute the metrics. Since we used the same evaluation code, we do not expect any deviations from previously reported results.

The evaluation distinguishes between five outcome types:

- **COR**: correct, exact match
- **INC**: incorrect match where the system and gold annotation disagree
- **PAR**: partial overlap between system and gold
- **MIS**: gold annotation missed by the system
- **SPU**: system prediction without a corresponding gold annotation

Based on these categories, we define *possible items* (POS) as

$$\text{POS} = \text{COR} + \text{INC} + \text{PAR} + \text{MIS},$$

representing all true positives and false negatives, and *actual items* (ACT) as

$$\text{ACT} = \text{COR} + \text{INC} + \text{PAR} + \text{SPU},$$

representing true positives and false positives.

Under *strict* matching, precision and recall are computed as

$$\text{Precision} = \frac{\text{COR}}{\text{ACT}} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{\text{COR}}{\text{POS}} = \frac{TP}{TP + FN}$$

For *partial* matching, overlapping annotations receive a weight of 0.5 to account for approximate agreement:

$$\text{Precision} = \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{ACT}}$$

$$\text{Recall} = \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{POS}}$$

Finally, the F1 score is reported for both settings as

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

[6] https://github.com/MantisAI/nervaluate

A.5 Mapping of categories to NER labels

Table 10 shows the mapping of each protagonist group class in the Moralization Corpus to the closest corresponding NER label.

protagonist group	NER label
Individuals	PER
Generic	PER
Institutions	ORG
Social Groups	MISC
Other	MISC

Table 10: Category mapping from protagonist groups labels (Moralization Corpus) to NER labels (flair).

A.6 Extended Evaluation Metrics for Protagonist Detection and Classification

Tables 11 and 12 present additional evaluation metrics, including precision and recall. For protagonist group classification, we report micro-F1 only (in the main section), as micro-precision, micro-recall, and micro-F1 are identical when exactly one label is predicted per instance.

	model name	experiment	f1		precision		recall	
			strict	partial	strict	partial	strict	partial
base LMs	bert-base-german-cased	fine_tuned	0.4325	**0.5288**	0.3917	**0.4790**	0.4826	0.5901
ner	flair-ner-german	base	0.1713	0.2521	0.2686	0.3952	0.1258	0.1851
		fine_tuned	0.4137	0.4413	**0.4515**	**0.4815**	0.3818	0.4072
	flair-ner-german-large	base	0.1695	0.2548	0.2559	0.3847	0.1267	0.1905
		fine_tuned	**0.4783**	**0.5375**	**0.4294**	**0.4826**	0.5397	0.6066
prompting	claude-3-5-haiku	pd_basic_0shot	0.3350	0.4667	0.2777	0.3870	0.4219	0.5879
		pd_cot_0shot	0.3586	0.5134	0.3071	0.4396	0.4309	0.6169
		pd_cot_10shot	0.3896	0.4955	0.3486	0.4433	0.4416	0.5616
		pd_cot_10shot_def	0.3896	0.5090	0.3364	0.4394	0.4630	0.6048
	gpt-5-mini	pd_basic_0shot	0.3504	0.3822	0.2450	0.2672	**0.6146**	**0.6704**
		pd_cot_0shot	**0.4789**	0.5163	0.3873	0.4176	**0.6271**	**0.6762**
		pd_cot_10shot	**0.4865**	0.5200	0.3939	0.4210	**0.6360**	**0.6798**
		pd_cot_10shot_def	**0.4882**	0.5211	0.3986	0.4255	**0.6298**	**0.6722**
human	human expert 2	annotation_context	0.4405	0.4835	0.3372	0.3702	0.6350	0.6971

Table 11: Additional results of protagonist detection experiments. Best performing models and scores that are not significantly different ($p \geq 0.05$) from the best-performing model are shown in **bold**. All values are micro-averaged.

	model name	experiment	context	f1	precision	recall	confidence
statistical baselines	ngram rule-based	fine_tuned		0.4368	0.4460	0.4280	-
	random forest	fine_tuned		0.4747	0.4830	0.4667	0.6160
base LMs	bert-base-german-cased	fine_tuned		0.5588	0.5086	**0.6201**	0.6697
		fine_tuned + masked context	✓	0.5895	0.5551	0.6285	0.7823
prompting	claude-3-5-haiku	prc_basic_0shot		0.4089	0.3619	0.4701	0.6666
		prc_cot_0shot		0.4443	0.3923	0.5122	0.6921
		prc_cot_10shot		0.2476	0.2337	0.2633	0.6554
		prc_cot_10shot_context	✓	0.2851	0.2606	0.3146	0.8167
		prc_cot_10shot_context+def	✓	0.2846	0.2663	0.3055	0.8030
	gpt-5-mini	prc_basic_0shot		0.3585	0.3139	0.4179	0.6735
		prc_cot_0shot		0.3432	0.3302	0.3572	0.4882
		prc_cot_10shot		0.3601	0.3524	0.3682	0.5074
		prc_cot_10shot_context	✓	0.6204	0.5862	0.6588	**0.8811**
		prc_cot_10shot_context+def	✓	**0.6426**	**0.6072**	**0.6824**	**0.8816**
human	human expert 1	annotation		0.5177	0.5328	0.5034	0.2978
		annotation_context	✓	0.6494	0.6135	0.6897	0.7853
	human expert 2	annotation		0.4752	0.4891	0.4621	0.5620
		annotation_context	✓	0.7203	0.7305	0.7103	0.8511

Table 12: Additional results of protagonist role classification experiments. Best performing models and scores that are not significantly different ($p \geq 0.05$) from the best-performing model are shown in **bold**. All values are micro-averaged.

A.7 Class-Specific Protagonist Classification Results

Table 13 and 14 show the class-specific protagonist classification results as an addition to Table 2a and 2b, respectively.

A.8 Statistical Significance

Figure 4 shows an example of a siginificance matrix (for the protagonist detection task, micro-F1); for all other experiments, metrices are provided in our project repository.[7]

The mapping between p-values and significance symbols in the Figures follows common reporting standards:

- $p < 0.001$: ******* (highly significant)
- $p < 0.01$: ****** (very significant)
- $p < 0.05$: ***** (significant)
- $p \geq 0.05$: *n.s.* (not significant)

A.9 Prompts

All prompts are available in our project repository.[7]

A.10 Semantic and Syntactic Categories for the Analysis of Context Effects on Protagonist Labels

To document which properties of protagonist phrases co-occur with changes in label assignment, we analyze a set of semantic (e.g., generic expressions such as *the people*, named entities, political reference, positive or negative sentiment) and syntactic features (e.g., phrase types and token length). The feature categories were derived from an initial bottom-up inspection of the data and informed by general considerations about context sensitivity.

For instance, longer noun phrases are expected to be more easily classifiable without context than pronouns; negatively connoted phrases (e.g., *war criminals*) are more likely to function as ADDRESSEES or MALEFICIARIES, whereas positively connoted phrases (e.g., *children*) are more often BENEFICIARIES; generic expressions (e.g., *the people*) can often be assigned without context and frequently appear as BENEFICIARIES; and references to political actors are typically classifiable as INSTITUTIONS and often function as ADDRESSEES or DEMANDERS (see Section 3).

This design allows us to systematically relate context effects to linguistically motivated properties of protagonist phrases. We label these categories manually for the subset of 50 instances (138 protagonist phrases); the complete distribution of features and label changes is reported in Table 15.

[7] https://github.com/anonymous-18122025/protagonist-detection-classification-moral

model name	experiment	context	Overall f1 exclud. Other	Individuals f1	Individuals conf	Institutions f1	Institutions conf	Generic f1	Generic conf	Social Groups f1	Social Groups conf	Other f1	Other conf
ngram rule-based	fine_tuned		0.5000	0.5564	-	0.5750	-	0.3203	-	0.4365	-	**0.0714**	-
random forest	fine_tuned		0.5269	0.5426	0.8539	0.5606	0.6019	0.4398	0.6348	0.5277	0.5677	**0.0000**	0.5116
bert-base-german-cased	fine_tuned		**0.7328**	**0.8593**	0.9399	**0.7931**	0.8669	**0.5290**	0.9138	0.6657	**0.9587**	**0.1224**	0.4691
	fine_tuned + masked context	✓	**0.7327**	0.8808	0.9401	**0.7945**	0.9114	**0.4907**	0.8207	**0.6777**	0.9037	**0.0370**	0.5214
claude-3-5-haiku	basic_0shot		0.6204	0.8202	0.9282	0.7248	0.9280	**0.4427**	0.8905	0.4621	0.8969	0.0845	0.5038
	cot_0shot		0.6924	**0.8519**	0.9309	0.7794	0.9350	**0.4854**	0.8559	0.6136	0.9066	**0.3061**	0.6189
	cot_10shot		0.6948	0.7903	0.9070	0.7735	0.9193	**0.5092**	0.8527	**0.6709**	0.8854	**0.2553**	0.4891
	cot_10shot_context	✓	0.7028	0.8450	0.9391	0.7610	0.9366	0.5256	0.8787	**0.6533**	0.8763	**0.2083**	0.3405
	cot_10shot_context+def	✓	0.7065	0.8588	0.9367	0.7416	0.9301	0.5236	0.8652	**0.6698**	0.8712	**0.1569**	0.3625
gpt-5-mini	basic_0shot		0.6756	**0.8153**	0.9212	0.7504	0.9306	0.3912	0.8968	**0.6304**	0.8863	**0.2524**	0.8069
	cot_0shot		0.7246	0.8502	0.9327	0.7919	0.9361	0.5070	0.8750	**0.6824**	0.8909	**0.2857**	0.7642
	cot_10shot		0.7242	**0.8396**	0.9143	0.7948	0.9296	**0.5013**	0.8026	**0.6909**	0.9060	**0.2093**	0.8244
	cot_10shot_context	✓	**0.7585**	**0.8654**	0.9588	**0.8290**	0.9497	**0.5580**	**0.9135**	**0.7068**	0.9325	**0.3636**	0.9031
	cot_10shot_context+def	✓	**0.7553**	**0.8848**	**0.9638**	**0.8124**	**0.9506**	**0.5435**	0.8989	**0.7092**	0.9408	**0.3014**	**0.8786**
human expert 1	annotation		0.6691	0.8475	0.8188	0.7857	0.6510	0.4615	0.7351	0.4865	0.7167	0.0000	0.0000
	annotation_context	✓	0.5802	0.8772	0.9667	0.5806	0.8821	0.3750	0.8980	0.5000	0.8174	0.6667	0.2000
human expert 2	annotation		0.5725	0.8421	0.9233	0.6465	0.8500	0.3380	0.7964	0.4865	0.7500	0.0000	0.0000
	annotation_context	✓	0.7126	0.9123	0.9867	0.7788	0.8896	0.4800	0.9400	0.5366	0.8813	0.6154	0.9000

Table 13: Protagonist group classification results broken down by role. Best performing models and scores that are not significantly different ($p \geq 0.05$) from the best-performing model are shown in bold. All values are micro-averaged. 'conf' denotes confidence.

model name	experiment	context	Overall f1 exclud. Unclear	Addressee f1	Addressee conf	Beneficiary f1	Beneficiary conf	Demander f1	Demander conf	Maleficiary f1	Maleficiary conf	Unclear f1	Unclear conf
ngram rule-based	fine_tuned		0.4797	**0.5067**	-	0.4904	-	0.5764	-	0.0000	-	**0.2164**	-
random forest	fine_tuned		0.5049	**0.4818**	0.6077	0.5756	0.6131	0.5707	0.6801	0.0000	0.7033	0.1202	0.4942
bert-base-german-cased	fine_tuned		0.5905	0.5795	0.6093	0.6632	0.7001	0.6619	0.6333	0.0000	0.0000	0.2609	0.3617
	fine_tuned + masked context	✓	0.6324	**0.6310**	0.7044	**0.6869**	0.8200	**0.7136**	0.8478	0.0000	0.0000	**0.3422**	0.5637
claude-3-5-haiku	basic_0shot		0.4478	0.4371	0.6978	0.5982	0.6842	0.3629	0.6982	0.3317	0.6978	0.1227	0.4706
	cot_0shot		0.4896	0.4936	0.7542	0.6657	0.7238	0.3733	0.7603	0.3192	0.7091	0.1194	0.4723
	cot_10shot		0.2623	0.3079	0.7489	0.2791	0.7044	0.2380	0.7361	0.1455	0.6980	0.1684	0.4426
	cot_10shot_context	✓	0.3024	0.3088	0.8506	0.3259	0.8277	0.3451	0.8438	0.2052	0.8356	0.0457	0.4087
	cot_10shot_context+def	✓	0.3004	0.2911	0.8309	0.3357	0.8122	0.3540	0.8317	0.1911	0.8146	0.0602	0.3568
gpt-5-mini	basic_0shot		0.4200	0.3832	0.5935	0.5589	0.6056	0.3286	0.5962	0.3191	0.6496	**0.1720**	0.6785
	cot_0shot		0.4043	0.3607	0.7155	0.4875	0.6848	0.4297	0.7020	0.2691	0.7000	**0.1839**	0.2633
	cot_10shot		0.4281	0.3927	0.7183	0.4613	0.7134	0.5035	0.6992	0.2949	0.6978	**0.1675**	0.2579
	cot_10shot_context	✓	**0.6503**	**0.6141**	0.8895	**0.7481**	**0.8967**	**0.6839**	0.8880	**0.4848**	0.8756	0.2591	**0.7125**
	cot_10shot_context+def	✓	**0.6684**	**0.6560**	**0.8916**	**0.7577**	0.8962	**0.7097**	**0.8917**	**0.4703**	**0.8745**	0.3263	0.6951
human expert 1	annotation		0.5290	0.5122	0.3229	0.6067	0.3707	0.7500	0.3652	0.2963	0.3688	0.0000	0.000
	annotation_context	✓	0.6833	0.7250	0.8780	0.6387	0.8563	0.7667	0.9185	0.5455	0.6455	0.2963	0.300
human expert 2	annotation		0.4855	0.4673	0.5600	0.4118	0.5600	0.6588	0.6192	0.0000	0.000	0.0000	0.000
	annotation_context	✓	0.7299	0.6972	0.8500	0.6988	0.8143	0.8710	0.9483	0.6000	0.8556	0.5000	0.600

Table 14: Protagonist role classification results broken down by role. Best performing models and scores that are not significantly different ($p \geq 0.05$) from the best-performing model are shown in bold. All values are micro-averaged. 'conf' denotes confidence.

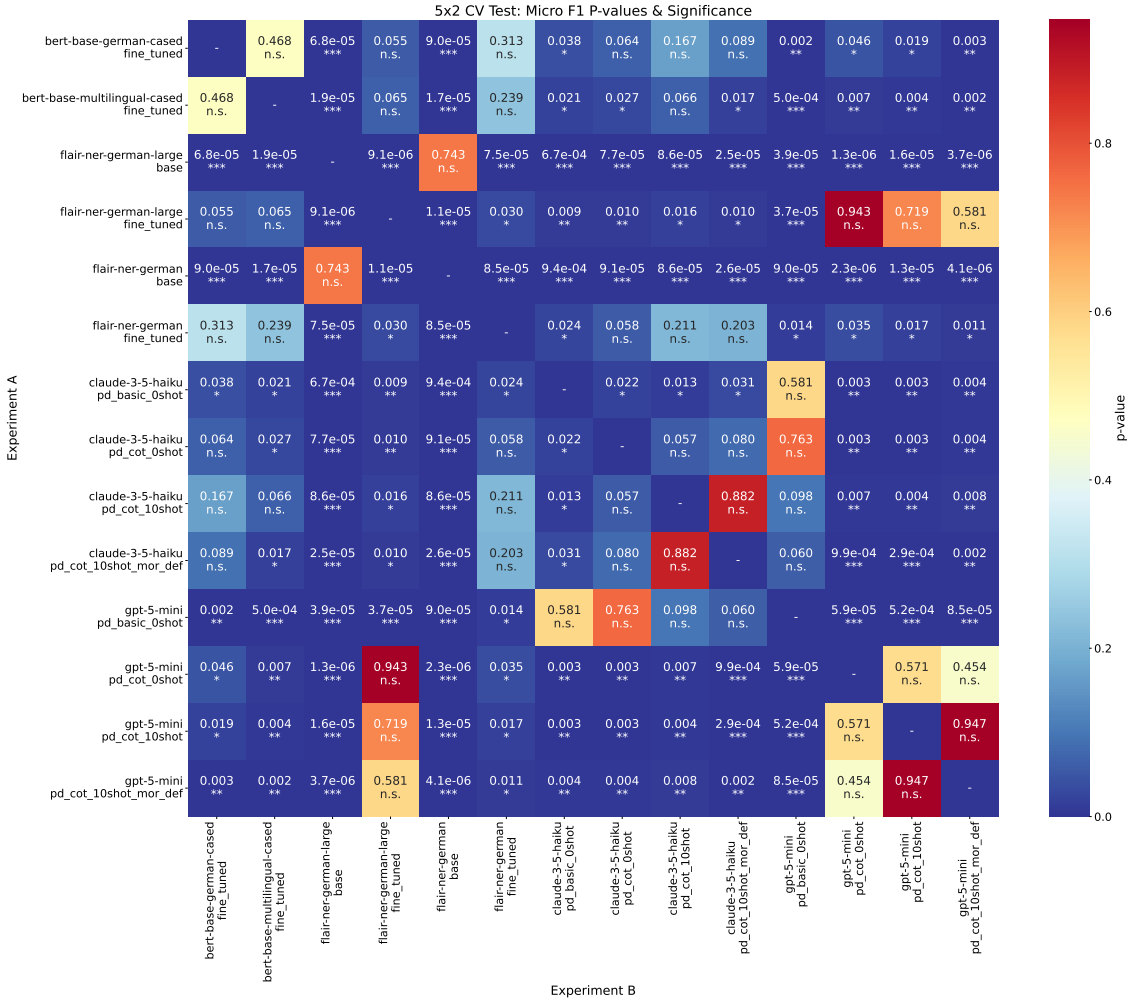


Figure 4: Statistical significance comparison of strict micro-F1 for Protagonist Detection and all model combinations.

	n	%	Generic Expressions	Named Entities	Reference to Politics	Positive Sentiment	Negative Sentiment
Groups - Humans with and without Context							
Labels left unchanged (Annotator 1 and 2 merged)	190	69.34	11.58	12.11	21.58	8.42	14.74
Labels that changed	84	30.66	2.40	3.60	15.50	11.90	16.70
Difference	**274**	**38.69**	**9.18**	**8.51**	**6.08**	**-3.48**	**-1.96**
Groups - CLAUDE with and without Context							
Labels left unchanged	119	86.86	10.08	9.24	18.49	10.92	15.13
Labels that changed	18	13.14	0.00	1.05	2.63	0.00	1.58
Difference	**137**	**73.72**	**10.08**	**8.19**	**15.86**	**10.92**	**13.55**
Groups - GPT with and without Context							
Labels left unchanged	123	89.78	3.25	2.44	4.07	0.00	1.63
Labels that changed	14	10.22	0.00	0.00	14.29	7.14	14.29
Difference	**137**	**-79.56**	**-3.25**	**-2.44**	**10.22**	**7.14**	**12.66**
Roles - Humans with and without Context							
Labels left unchanged (Annotaor 1 and 2 merged)	130	47.45	9.23	6.92	15.38	10.77	10.77
Labels that changed	144	52.55	8.33	11.81	23.61	8.33	19.44
Difference	**274**	**-5.11**	**0.90**	**-4.88**	**-8.23**	**2.44**	**-8.68**
Roles - CLAUDE with and without Context							
Labels left unchanged	112	81.75	8.04	8.04	17.86	10.71	16.07
Labels that changed	25	18.25	28.00	4.00	12.00	9.12	0.00
Difference	**137**	**63.50**	**-19.96**	**4.04**	**5.86**	**1.59**	**16.07**
Roles - GPT with and without Context							
Labels left unchanged	42	30.66	9.52	9.52	19.05	11.90	21.43
Labels that changed	95	69.34	8.42	9.47	20.00	8.42	12.63
Difference	**137**	**-38.69**	**1.10**	**0.05**	**-0.95**	**3.48**	**8.80**

(a) Semantic features.

	NP	PronNP	PP	Noun	Pronoun	Number of tokens
Groups - Humans with and without Context						
Labels left unchanged (Annotator 1 and 2 merged)	71.58	3.13	4.25	6.84	14.21	2.15
Labels that changed	53.66	2.44	2.44	6.10	35.37	2.21
Difference	**17.92**	**0.69**	**1.81**	**0.74**	**-21.16**	**-0.06**
Groups - CLAUDE with and without Context						
Labels left unchanged	65.62	1.91	1.99	10.12	20.36	2.15
Labels that changed	64.71	5.88	5.88	0.00	23.53	2.39
Difference	**0.91**	**-3.97**	**-3.89**	**10.12**	**-3.17**	**-0.24**
Groups - GPT with and without Context						
Labels left unchanged	66.39	1.68	1.68	9.24	21.01	2.12
Labels that changed	60.00	0.00	13.33	6.67	20.00	2.71
Difference	**-6.39**	**-1.68**	**11.65**	**-2.57**	**-1.01**	**0.59**
Roles - Humans with and without Context						
Labels left unchanged (Annotaor 1 and 2 merged)	56.15	3.37	5.08	7.39	28.01	2.16
Labels that changed	74.31	3.08	1.39	5.95	15.28	2.20
Difference	**-18.16**	**0.29**	**3.69**	**1.44**	**12.73**	**-0.04**
Roles - CLAUDE with and without Context						
Labels left unchanged	66.07	1.89	1.89	8.04	22.12	2.17
Labels that changed	64.00	6.53	9.47	0.00	20.00	2.29
Difference	**2.07**	**-4.64**	**-7.58**	**8.04**	**2.12**	**-0.12**
Roles - GPT with and without Context						
Labels left unchanged	66.29	4.38	0.00	5.90	23.43	2.40
Labels that changed	66.32	2.05	3.16	7.37	21.10	2.09
Difference	**-0.03**	**2.33**	**-3.16**	**-1.47**	**2.33**	**0.31**

(b) Syntactic features.

Table 15: Context effects on protagonist label changes by semantic and syntactic categories. Values in % (except for *n*).