Identifying Rhetorical Roles from Dutch Judicial Decisions: Relevance Detection, Role Classification, and Bayesian Structural Updating

Anonymous ACL submission

Abstract

Classifying the material facts, submissions, judicial reasoning, and the court's decision (rhetorical roles) in case law is a prerequisite for high-quality of virtually all NLP-in-law tasks, including legal information retrieval, summarisation, and LLM-based analysis. A problem that has been left unaddressed is how to reliably classify these rhetorical roles if the roles are intertwined and hence change sentence-by-sentence: one sentence discusses a party's claim, the next a material fact, and the next concerns judicial reasoning. This paper presents a study of sentence-level rhetorical role classification for Dutch judicial decisions. We introduce a manually annotated dataset labelled across four legally meaningful categories — material facts, submissions, judicial reasoning, and court's decision — alongside non-annotated content. We formalise and compare different classification strategies. We also test a Bayesian structural update that incorporates section header priors into classifier confidence. The empirical results across all strategies are reported, and the implications for legal NLP discussed. Our results show that explicitly separating relevance from role classification and incorporating structural signals yields consistent performance gains on this challenging, sparsely annotated task

1 Introduction

Those who have used automated AI tools to retrieve or search a judicial decision (Ma et al., 2024; Zhong and Litman, 2022) and found that the results were almost (but not quite) right have encountered the problem central in this study: Nearly every computational task in the legal domain depends on knowing what role a sentence plays (Walker et al., 2019; Bhattacharya et al., 2019; Malik et al., 2022; Nigam et al., 2025). Retrieval systems that search specifically within a factual background (Lou et al., 2021; Shulayeva et al., 2017) or reasoning section (Csányi et al., 2025) return more targeted results than those treating decisions as undifferentiated text. Summaries that distinguish facts from reasoning and conclusions outperform those that do not (Zhong and Litman, 2022; Kalamkar et al., 2022). For LLM-based systems (Magesh et al., 2025; Dahl et al., 2024), whether fine-tuned or prompted (Chalkidis et al., 2020; Deroy et al., 2024), absent segmentation means the model cannot reliably distinguish a court's judgment from its reasoning or procedural context - a bottleneck restricting everything built downstream (Ariai and Demartini, 2024; Kalamkar et al., 2022).

The classification of rhetorical roles in judicial decisions is often treated as tractable (Walker et al., 2019; Bhattacharya et al., 2019; Savelka and Ashley, 2018; Malik et al., 2022; Kalamkar et al., 2022; Nigam et al., 2025; Santosh et al., 2023, 2024; Csányi et al., 2025; Darji et al., 2026; Lou et al., 2021; Xu et al., 2021). In practice it is not. As Figure 1 illustrates, the categories of interest do not cluster neatly within document sections. In fact, the facts are often intertwined with what parties claim or submit and what the court reasons and decides. Sentences with a certain rhetorical role might be preceded and succeeded by sentences with a different role. How to disentangle the rhetorical roles when taking the context into consideration offers limited value?

This paper studies sentence-level extraction of legally relevant information from Dutch judicial decisions across four core categories: *material facts* (*materiële feiten*), *submissions* (*proceshandelingen*), *judicial reasoning* (*beoordeling*), and *court's decision* (*beslissing*), alongside non-annotated content. We make three contributions: (1) a sentence-level annotated dataset of Dutch judicial decisions labelled for rhetorical role and legal relevance; (2) a formalisation and comparison of classification strategies varying along these two dimensions; and (3) empirical results across strategies with implica-

ANNOTATED LEGAL EXCERPT

Executory sale of her home would, under the given circumstances, constitute abuse of enforcement authority. procedural actions The court rules as follows. It follows from the agreement between the parties that [appellant] shall repay the borrowed amount when she is able to do so. material facts Pursuant to Article 7A:1798 of the Civil Code, the court must in such a case determine the time of repayment "according to the circumstances". On the basis of the mutual submissions of the parties it may be assumed that [appellant] will no longer receive income from her business. material facts It could have been expected of [appellant] that she would also provide insight into her financial position. assessment

ORIGINAL — DUTCH

Executoriale verkoop van haar woning zou in de gegeven omstandigheden misbruik van executiebevoegdheid opleveren. Het hof oordeelt als volgt. Uit de overeenkomst tussen partijen volgt dat [appellante] het geleende bedrag zal terugbetalen wanneer zij daartoe in staat zal zijn. Ingevolge artikel 7A:1798 BW dient de rechter in zodanig geval de tijd der teruggave te bepalen "naar gelang der omstandigheden". Op basis van de wederzijdse stellingen van partijen kan worden aangenomen dat [appellante] geen inkomsten meer zal ontvangen uit haar onderneming. Van [appellante] had verwacht mogen worden dat zij ook inzicht zou geven in haar vermogenspositie.

Figure 1: An annotated excerpt from the assessment section (beoordeling) of a Dutch judicial decision, illustrating how material facts, procedural actions, and court assessments interleave within a single passage along with unlabelled (not related) content.

tions for legal NLP.

2 Related Work

Malik et al. (2022) established rhetorical role classification as a well-defined NLP task with a large-scale annotated corpus of Indian legal judgments across 13 roles, serving as a reference point for subsequent work. Kalamkar et al. (2022) extended this with a corpus targeting summarisation and judgment prediction. Nigam et al. (2025) scaled these efforts with LegalSeg (7,000+ documents, 1.4M sentences), finding that models incorporating broader context consistently outperform sentence-level-only approaches. At SemEval-2023, Tandon and Chatterjee (2023) combined topic modelling with RoBERTa and a CRF layer, achieving micro-F1 of 0.798. A recurring finding is that classification improves when structural cues are incorporated. Xu et al. (2021) showed that sentence position combined with domain-specific embeddings improves classification into issues, conclusions, and reasons — motivating our Bayesian structural update, which treats section headers as a probabilistic signal rather than a hard override. Santosh et al. (2023) reformulated the task at span level using a semi-Markov CRF, improving over sentence-level baselines, with a span-to-sentence mapping parallel to our annotation design. Santosh et al. (2024) showed that semantically similar instances at inference time yield consistent macro-F1 gains. Walker et al. (2019) found that rule-based approaches plateau on larger datasets where learned models continue to improve, informing our fully data-driven pipeline. The work most directly comparable to ours is Csányi et al. (2025), who studied sentence-level rhetorical role labelling for Hungarian judicial decisions with iterative human annotation, evaluating a range of neural architectures and embedding strategies. Our work differs in language (Dutch), in explicitly distinguishing relevance detection (i.e., first filtering sentences that should receive *a* (any) label) from role classification, and in introducing a Bayesian structural update component. Lou et al. (2021) offer a complementary perspective on French decisions, applying paragraph segmentation to extract case-relevant facts and distinguishing what a sentence says from the function it serves — a distinction central to our annotation scheme. For Dutch legal text specifically, van Drie et al. (2023) introduced a sentence-level annotated dataset of legislative text labelled for semantic roles; though focused on statutory rather than judicial text, the annotation challenges they identify — stylistic variability, domain vocabulary, and sparse labelling — closely reflect those encountered in our study.

3 Dataset

3.1 Dataset Origin and Scope

The raw data used in this study consists of 102 Dutch judicial decisions collected from Rechtspraak.nl, the official public portal of the Dutch judiciary. All documents were obtained in their original XML format. The section headers present

in the XML was preserved for each document. The dataset covers three major areas of Dutch law: civil law (*Burgerlijk Wetboek*), criminal law (*Wetboek van Strafrecht*), and administrative law (*Algemene Wet Bestuursrecht*).

Cases were randomly sampled from a curated pool of decisions restricted to 2015–2021 and selected from court codes manually curated by a legal expert to ensure representation across different court types and jurisdictions, with an initial target of 34 cases per domain. Following preprocessing, two cases were excluded due to incorrect segmentation, resulting in a final dataset of 100 judicial decisions distributed across 33 civil, 34 criminal, and 33 administrative cases, drawn from 16 distinct courts. Each decision retains its European Case Law Identifier (ECLI), supporting full traceability and case-level data splitting.

3.2 Annotation

The annotation scheme is built around four substantive labels — *materiële feiten*, *beoordeling*, *beslissing*, and *proceshandelingen* — which constitute the core argumentative and decisional structure of a judicial decision: *materiële feiten* captures facts directly forming the factual basis of a claim or defence; *proceshandelingen* covers procedural steps and formal acts, including parties' positions, claims, and submitted documents; *beoordeling* encompasses the legal reasoning in which the judge explains how and why they reached a decision; and *beslissing* records the final ruling. A fifth label, `None`, was assigned to all sentences (while creating the dataset) not covered by any substantive label. The full annotation scheme distinguishes eight categories; the remaining four served to inform the annotators as to what should *not* be annotated and why (see Appendix A for full details). An important feature of the annotation guidelines is that category assignment is context-dependent: a sentence ordinarily classified as *proceshandelingen* may be reclassified as *materiële feiten* if it directly informs the substantive assessment of the dispute. For example, "The summons was not received" would normally indicate a procedural irregularity, but in a case concerning a violation of the right to a defence it constitutes a material fact. Similarly, "The Amsterdam Court of Appeal convicted the defendant" may describe a procedural event in an appeal, but where the defendant's criminal liability is at issue it bears directly on the substantive assessment.

The dataset was annotated by one law professor and one PhD student in law, following detailed written guidelines. Prior to full-scale annotation, both annotators independently labelled a shared subset across eleven rounds, refining the guidelines iteratively. Inter-annotator agreement was then assessed on 20 jointly labelled decisions, yielding a Cohen's kappa of 0.344 and observable agreement of 0.695, and that was after the eleven training rounds. The relatively low kappa partly reflects high chance agreement under an imbalanced label distribution. Remaining disagreements concerned boundary segmentation. Annotators consistently agreed on label identity but differed on sentence boundaries. All disagreements were resolved through a structured consensus discussion, producing a fully adjudicated gold standard. The full dataset of 102 cases was then divided evenly, with each annotator independently labelling 51 decisions.

3.3 Preprocessing and Sentence Segmentation

Decisions were extracted from their XML representation and converted to plain text while preserving section boundaries and their corresponding headers. Sentence segmentation was performed using a custom rule-based sentence splitter developed specifically for Dutch legal texts, incorporating domain specific heuristics to handle legal writing conventions such as abbreviations, enumerations, and irregular formatting. Despite these tailored rules, perfect segmentation could not be achieved given the inherent complexity of judicial texts. Sentences resulting from clear segmentation errors were removed prior to experimentation, totalling 70 sentences, less than 1% of the original total. Full details of the preprocessing pipeline are provided in Appendix A. The sentence splitter is released together with the dataset.

3.4 Header Information and Mapping

Dutch judicial decisions are typically divided into titled sections, and these section headers provide useful cues about the local role of the text. Each sentence was matched to the header of the section from which it originated. Because raw section titles vary substantially across decisions, headers were normalized into broader functional groups using manually constructed mappings presented in Appendix A. The final retained header groups were Context, Feiten (Facts), Beoordeling (Reasoning), Beslissing (Decision), and Proceshandelingen partijen (Party procedural actions). Three broader categories were excluded because they fell outside the

scope of the annotation task: Wettelijk kader (Legal framework), Proceskosten (Legal costs), and Procesverloop (Procedural history). A residual Other category was used for headers that could not be mapped reliably. The 70 sentences assigned to this category correspond to the segmentation errors removed during preprocessing, as described above.

3.5 Dataset Overview

The final dataset contains 8,391 sentences from 100 judicial decisions. Table 1 presents the label distribution, which reflects the natural structure of judicial decisions: None accounts for 36.5% of sentences, while *beslissing* represents only 2.9%, consistent with the conciseness of final rulings. All splits were performed at the case level to prevent information leakage, yielding 5,608 training sentences (66.8%), 1,281 validation sentences (15.3%), and 1,502 test sentences (17.9%).

Table 1: Label distribution in the annotated dataset.

Label	Count	Proportion
None	3,060	36.5%
beoordeling	2,293	27.3%
materiële feiten	1,529	18.2%
proceshandelingen	1,263	15.1%
beslissing	246	2.9%
Total	8,391	100.0%

The dataset constitutes a core contribution of this study and will be made publicly available, including annotated sentence-level data, ECLI identifiers, and structural metadata extracted from the original XML documents.

4 Methodology

The methodology is organized around two dimensions: whether relevance detection and legal-role assignment are handled jointly or separately, and whether incorporating structural information, specifically section headers, improves the classification. The resulting strategies are model-agnostic and are evaluated using classical supervised baselines, fine-tuned encoder models, and prompt-based large language models.

4.1 Classification Task Formulation

The annotation scheme defined a five-label sentence classification task consisting of four substantive legal-role categories and the None category. Since not every sentence contributed substantively to the legal analysis, we distinguished between **relevance detection** — identifying whether a sentence carried any legal role at all — and **role assignment** — determining which specific legal role it performed.

Four supervised classification formulations were evaluated. The first, referred to as the **Five-way Classification** formulation, predicted one of the five final labels in a single step: None, *materiële feiten*, *proceshandelingen*, *beoordeling*, or *beslissing*, treating relevance detection and role assignment as a single classification problem. The second, referred to as **Relevance Detection**, was binary relevance detection, where the model predicted whether a sentence was None or belonged to one of the four substantive legal-role categories, tackling the problem of identifying "relevant" sentences that can later be classified under of the four labels. The third, referred to as **Four-Way Role Classification**, excluded None sentences and was trained and evaluated only on gold legally labelled sentences. It therefore measured how well a model distinguished between legal roles once relevance had already been established. The fourth, referred to as **One-vs-Rest Role Detection (OvR)**, trained separate binary classifiers for each substantive legal-role category. In contrast to the other formulations, this allowed a single sentence to carry multiple labels.

4.2 Model Families

TThree model families were evaluated. Logistic Regression with TF-IDF features (Pedregosa et al., 2011) served as a classical supervised baseline, providing a lexical reference point against which contextual models can be compared. We fine-tuned two Dutch encoder-based language models — RobBERT (Delobelle et al., 2020) and BERTje (de Vries et al., 2019) — selected for their strong performance on Dutch NLP tasks and their suitability for sentence-level classification. Each model was trained separately for each task formulation using the same frozen train, validation, and test splits. Finally, we evaluated four open-source instruction-tuned large language models in a prompting-based setup — Gemma-3-12B-IT (Gemma Team, Google DeepMind, 2025), Reynaerde-7B-Chat (Van den Avenne, 2024), Qwen2.5-7B-Instruct (Qwen Team, 2024), and Llama-3-8B-Instruct (Grattafiori et al., 2024) — to assess whether general-purpose Large Language Models can handle this task without supervised

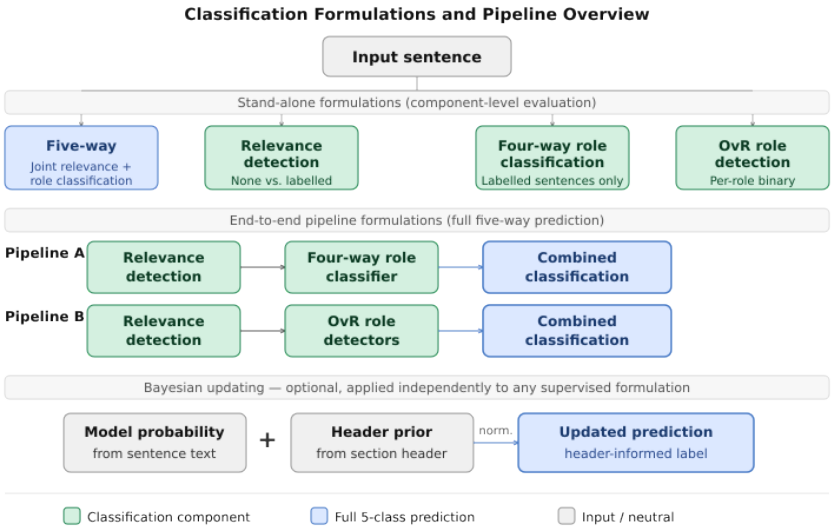


Figure 2: Overview of classification formulations, end-to-end pipeline combinations, and Bayesian updating. Teal boxes indicate components; purple boxes indicate final five-way predictions.

fine-tuning. The details for prompt settings and the prompts used can be found in Appendix B.

4.3 End-to-End Pipeline Formulations

Figure 2 illustrates how the component-level formulations are combined into two end-to-end pipelines. Both pipelines use probabilistic chaining, as explained below, preserving uncertainty from Stage 1 rather than applying a hard relevance gate.

Pipeline A combines Relevance Detection with the Four-Way Role Classifier:

$$P(\text{None} \mid s) = P(z = 0 \mid s),$$

$$P(y \mid s) = P(z = 1 \mid s) \cdot P(y \mid s, z = 1).$$

Pipeline B combines Relevance Detection with the One-vs-Rest detectors. Since the OvR classifiers were trained independently, their positive-class scores are normalized across the four substantive labels before being weighted by the Stage 1 probability of legal relevance.

4.4 Header-Based Bayesian Updating

Each sentence was associated with a section header preserved from the original XML structure. For each header group, we estimated from the training data how often each label occurred under that header, yielding a header-based prior $P(y \mid h)$. At inference time, the text-model probability was combined with this prior:

$$\log P(y \mid s, h) \propto \log P(y \mid s) + \lambda \log P(y \mid h),$$

where λ controls the strength of the header signal. We used $\lambda = 1$, to make sure the header signals do not overpower the classification model. The resulting scores were normalized and the highest-scoring label was selected. Bayesian updating was applied only to supervised models with probabilistic outputs.

5 Experiments & Results

5.1 Overall Results

Table 2 reports the macro F1 scores of the supervised models before and after Bayesian updating. Macro F1 is used as the main comparison metric because the label distribution is imbalanced and the task requires performance across all legal-role classes. Prompting-based LLMs were also evaluated, but they performed substantially below the supervised models; the LLM results are reported in Table 6 (Appendix). The full supervised results, including accuracy, precision, recall, and weighted F1, are reported in Table 5 (Appendix).

The Dutch encoder models generally outperform the Logistic Regression baseline. In the direct five-way setting, RobBERT obtains the strongest result before Bayesian updating, with a macro F1 of 0.6680. BERTje follows closely with 0.6561, while Logistic Regression remains lower at 0.5726. The same pattern appears in the relevance-detection and role-classification settings, where the contextual models are consistently stronger than the lexical

Table 2: Macro F1 performance of supervised models before and after Bayesian updating. Values report macro F1. Δ denotes the change after Bayesian updating. Stage 1 is evaluated as a binary classification task distinguishing unlabelled sentences from legally labelled sentences. Stage 2 and OvR are evaluated only on gold legally labelled sentences. Pipeline A combines Stage 1 with the Stage 2 four-way classifier, while Pipeline B combines Stage 1 with normalized OvR probabilities. Best macro F1 scores within each setting are shown in bold. Full accuracy, precision, recall, macro F1, and weighted F1 results are reported in Appendix Table 5

Model	Setting	Evaluation subset	Support	Before Bayes	After Bayes	Δ
Logistic Regression	5-way	All sentences	1502	0.5726	0.4439	-0.1287
	Relevance detector	All sentences, binary	1502	0.6704	0.6116	-0.0588
	4-way	Legal-labelled only	961	0.6970	0.7429	+0.0459
	OvR mean	Legal-labelled only	961	0.7866	0.8116	+0.0250
	Pipeline A	All sentences	1502	0.4949	0.4926	-0.0023
	Pipeline B	All sentences	1502	0.4665	0.4953	+0.0288
BERTje	5-way	All sentences	1502	0.6561	0.6549	-0.0012
	Relevance detector	All sentences, binary	1502	**0.7592**	0.7556	-0.0036
	4-way	Legal-labelled only	961	0.7272	0.7965	+0.0693
	OvR mean	Legal-labelled only	961	0.8118	0.8270	+0.0152
	Pipeline A	All sentences	1502	0.6164	**0.6679**	+0.0515
	Pipeline B	All sentences	1502	0.6212	**0.6717**	+0.0505
RobBERT	5-way	All sentences	1502	**0.6680**	0.6071	-0.0609
	Relevance detector	All sentences, binary	1502	0.7562	0.6944	-0.0618
	4-way	Legal-labelled only	961	0.7526	**0.8064**	+0.0538
	OvR mean	Legal-labelled only	961	0.8328	**0.8351**	+0.0023
	Pipeline A	All sentences	1502	0.6513	0.6539	+0.0026
	Pipeline B	All sentences	1502	0.6604	0.6711	+0.0107

baseline. Sentence-level legal-role classification therefore benefits from contextual representations rather than relying only on surface-level lexical features.

Bayesian updating has a task-dependent effect. It does not improve the direct five-way setting or Stage 1 relevance detection. In these settings, the model still has to separate legally labelled sentences from the broad *None* category, and the section-header priors do not provide a reliable signal for that decision. The clearest gains appear after this broad filtering problem has been removed. In Stage 2 four-way classification, all supervised models improve after Bayesian updating. RobBERT obtains the strongest Stage 2 result, increasing from 0.7526 to 0.8064 macro F1. The OvR setting shows a smaller improvement, with RobBERT increasing from 0.8328 to 0.8351.

The pipeline results capture a different trade-off. The pipelines are not only intended to maximize direct five-way classification accuracy; they represent a downstream-oriented formulation in which relevance detection and role assignment are separated. Among the pipeline variants, BERTje Pipeline B with Bayesian updating obtains the strongest macro F1 of 0.6717, slightly above the direct five-way RobBERT score of 0.6680. The pipeline results therefore remain competitive with direct classification, even though they separate relevance detection from role assignment.

5.2 Class-Level Analysis

Given the overall model comparison, the class-level analysis focuses on the RobBERT-based components. RobBERT obtains the strongest results for the main role-classification settings, namely four-way classification and OvR role detection. Although BERTje performs slightly better in Stage 1 relevance detection, the difference from RobBERT is small. We therefore inspect the three RobBERT components, Stage 2 role classification, and OvR role detection.

Table 3 reports the class-level results for the RobBERT 4-way and OvR role-classification models. Across the role labels, *beslissing* is the most stable class, with F1 scores above 0.89 in all four settings. *Beoordeling* also performs consistently across the Stage 2 and OvR formulations. The lower scores are concentrated in *materiële feiten* and *proceshandelingen*, which show greater variation across model formulation and Bayesian updating.

Table 3: RobBERT role-classification results for 4-way and one-vs-rest models. S2 refers to the four-way classifier evaluated only on gold legally labelled sentences. OvR refers to one-vs-rest models and reports the positive class for each legal-role detector. S2+B and OvR+B refer to the same models after Bayesian updating with section-header priors.

Class	Support	Precision				Recall				F1-score			
		S2	S2+B	OvR	OvR+B	S2	S2+B	OvR	OvR+B	S2	S2+B	OvR	OvR+B
beoordeling	389	0.736	0.732	0.762	0.842	0.833	0.872	0.805	0.782	0.782	0.796	0.783	0.811
beslissing	69	0.968	0.983	0.969	0.983	0.870	0.826	0.899	0.841	0.916	0.898	0.932	0.906
materiële feiten	297	0.772	0.853	0.772	0.967	0.650	0.781	0.673	0.492	0.706	0.816	0.719	0.652
proceshandelingen	206	0.603	0.798	0.679	0.891	0.612	0.651	0.534	0.476	0.607	0.717	0.598	0.620
Macro avg	961	0.770	0.841	–	–	0.741	0.782	–	–	0.753	0.806	–	–

In the four-way model, Bayesian updating improves all four legal-role labels. The largest gains appear for *materiële feiten*, where F1 increases from 0.706 to 0.816, and *proceshandelingen*, where F1 increases from 0.607 to 0.717. The improvement is smaller for *beoordeling*, from 0.782 to 0.796, while *beslissing* remains high but decreases slightly from 0.916 to 0.898.

The OvR models show a different pattern. After Bayesian updating, *beoordeling* improves from 0.783 to 0.811 F1, while *beslissing* remains high, decreasing from 0.932 to 0.906. For *materiële feiten* and *proceshandelingen*, Bayesian updating increases precision but lowers recall. For example, *materiële feiten* precision increases from 0.772 to 0.967, while recall decreases from 0.673 to 0.492. A similar precision–recall shift appears for *proceshandelingen*. These class-level results show that Bayesian updating affects the four-way and OvR formulations differently, with the four-way model showing more balanced F1 gains and the OvR models showing stronger changes in precision and recall.

Table 4: RobBERT relevance-detection class-level results before and after Bayesian updating. Labelled includes all sentences that received one of the fourlabels in the study.

Class	Support	Precision		Recall		F1-score	
		Before	After	Before	After	Before	After
None	541	0.749	0.775	0.608	0.440	0.671	0.561
Labelled	961	0.801	0.746	0.886	0.928	0.841	0.828
Macro avg	1502	0.775	0.761	0.747	0.684	0.756	0.694
Weighted avg	1502	0.782	0.757	0.786	0.752	0.780	0.732

Table 4 reports the class-level results for Stage 1 relevance detection. Before Bayesian updating, RobBERT obtains a higher F1 score for legally labelled sentences than for *None* sentences, with scores of 0.841 and 0.671 respectively. After Bayesian updating, recall for labelled sentences increases from 0.886 to 0.928, while precision decreases from 0.801 to 0.746. For the *None* class, F1 decreases from 0.671 to 0.561.

6 Discussion

The central point raised at the start of this paper is that judicial decisions cannot simply be treated as undifferentiated text, whereas for many legal NLP tasks it matters whether a sentence states a material fact, reports a procedural act, forms part of the court's reasoning, or expresses the final decision. Parsing different sections might prove insufficient to capture these roles, as the roles can (and often do) change sentence-by-sentence. This makes the classification task difficult: one must separate legally meaningful sentences from unlabelled material *and* distinguish between legal roles that are not always cleanly separated.

What is the best approach depends on what the system is meant to do downstream. Some uses are more **coverage-oriented**, favouring high recall to avoid missing relevant sentences. This approach can be useful for legal search, candidate passage selection, or the first stage of a summarisation pipeline, where some noise may be acceptable if important material is preserved. For example, a relevance filter that is too strict early on may cause important facts or decisions to never reach the summarisation model at all. Other uses are more **precision-oriented**. There, tolerance for noise becomes a liability. This applies to, for instance, provision retrieval based on material facts. In the last setting, the input should ideally contain only the factual basis of the case, as information about the reasoning might provide cues regarding the provision that is searched. High precision for material facts may therefore be more important for these types of tasks than broad coverage. The model formulations tested in this study can be understood as occupying different positions along this precision-recall spectrum.

The **direct five-way model** treats relevance detection and role assignment as a single end-to-end classification task, assigning one label to every sentence including `None`. It performs well on `None`, `beoordeling`, and `beslissing`, but has more difficulty with `materiële feiten` and `proceshandelingen`, which are more often confused with `None`, `beoordeling`, or each other (Figure 3). The five-way model is useful as a general-purpose document-labelling baseline, but its errors also show that a single end-to-end classifier may not be sufficient for all downstream tasks.

The **relevance-detection model** is the most coverage-oriented formulation. Rather than assigning a specific legal role, it asks only whether a sentence is legally meaningful at all. The model achieves high recall for legally labelled sentences, meaning relatively few relevant sentences are missed, at the cost of some `None` sentences being passed forward (Table 4; Figure 4). Bayesian updating shifts this balance, increasing recall but reducing precision. The without-Bayesian-update model is more balanced, while the with-Bayesian-update model is more coverage-oriented for legally labelled material. The latter suits coverage-oriented tasks such as legal search or the first stage of a summarisation pipeline, whereas tasks that use retrieved sentences directly may require stricter filtering.

The **four-way model** operates under a different assumption: that relevance has already been established, and the task is only to distinguish between substantive legal roles. `Beslissing` performs best, as court decisions tend to be linguistically distinctive and structurally predictable. `Materiële feiten` and `proceshandelingen` are harder — they are contextually defined and more easily confused with one another, as reflected in Figure 5. Bayesian updating has the largest effect here, with macro F1 improving from 0.753 to 0.806 (Table 3), suggesting that section-header information does substantial work even for categories that are not always confined to clearly defined sections of a judgment. In settings where role assignment matters, such as structured case summaries, the header-informed four-way model offers a meaningful precision gain.

The **one-vs-rest models** treat each legal role as a separate binary detection problem, making them the most precision-oriented formulation. Bayesian updating pushes some detectors toward an extreme position on the spectrum: for `materiële feiten`, precision reaches 0.967 after updating, but recall drops to 0.492, meaning the model retrieves only the highest-confidence instances while missing nearly half of the true ones. The effect is not uniform — `beslissing` remains stable with high precision and recall after updating, while the tradeoff is most pronounced for `materiële feiten` and `proceshandelingen` (Table 3; Figures 6 and 7). For a provision retrieval system, this level of precision may be exactly what is needed, even at the cost of coverage. For training-data construction or structured case summaries, the without-Bayesian update detector offers a more balanced alternative.

The pipeline formulation adds a final qualification. Although the staged setup does not clearly outperform the direct five-way model, its performance remains close enough to make the comparison meaningful. This suggests that the pipeline should not be understood only as a way to improve aggregate classification scores. Its main value is methodological. This is useful because downstream systems may require different levels of strictness at each stage. If these components are optimized according to the needs of a specific downstream task, the pipeline formulation may therefore be more useful than a single direct five-way classifier, even when their overall scores are similar.

7 Conclusion and Future Work

This study shows that sentence-level legal-role classification is best understood as a set of modeling choices shaped by the legal and argumentative context, rather than as a search for one universally best classifier. The results reveal that judicial decisions cannot be treated as undifferentiated text: rhetorical roles are interleaved, context-dependent and only partially aligned with the document structure.

The experiments further show that different classification formulations explored, occupy different positions along a precision-recall spectrum. Separating relevance detection from role assignment yields more flexible behavior, while Bayesian updating shows that information about the headings a sentence belongs to becomes more useful once irrelevant material has been filtered out.

This work contributes a Dutch sentence-level annotated dataset, a comparative modeling framework and above all evidence that structurally informed formulations provide a stronger foundation for different legal NLP systems. Future work can explore the impact of these formulations to different tasks.

8 Limitations

The dataset covers 100 Dutch judicial decisions across three legal domains. While this is sufficient to establish proof of concept and draw meaningful comparisons across model formulations, it does not capture the full variation present in the broader Dutch judiciary. Extending the dataset and testing across jurisdictions is a natural next step. Inter-annotator agreement was moderate, with a Cohen's kappa of 0.344. This partly reflects the genuine difficulty of the task — as the annotation examples in Appendix A illustrate, many sentences are ambiguously situated between categories, and disagreements often concern boundary cases rather than clear labelling errors. The iterative refinement process across eleven rounds and the detailed annotation guidelines were designed precisely to manage this complexity. The models classify sentences in isolation, without access to surrounding context. This is a known limitation of sentence-level approaches, and one that the results already account for — the Bayesian structural update can be understood as a lightweight way of reintroducing some of that contextual signal through section headers. Performance on context-dependent categories like materiële feiten and proceshandelingen would likely improve further with sequence-level modelling. Finally, all evaluations are intrinsic. Whether rhetorical role classification tested in this study improves downstream tasks such as retrieval or summarisation remains an open question, and one this study is not positioned to answer — but the formulations developed here are designed with exactly those applications in mind.

References

Farid Ariai and Gianluca Demartini. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in Indian legal judgments. In *Legal Knowledge and Information Systems: JURIX 2019*, pages 3–12. IOS Press.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Gergely Márk Csányi, István Üveges, Dorina Lakatos, Dóra Ripszám, Kornélia Kozák, Dániel Nagy, and János Pál Vadász. 2025. Sentence-level rhetorical role labeling in judicial decisions. *Big Data and Cognitive Computing*, 9(12):315.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Harshil Darji, Martin Heckelmann, Christina Kratsch, and Gerard de Melo. 2026. Segmentation and processing of German court decisions from open legal data. In *Legal Knowledge and Information Systems: JURIX 2026*. IOS Press.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024. Applicability of large language models and generative models for legal case judgement summarization. *arXiv preprint arXiv:2407.12848*.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Andrés Lou, Olivier Salaün, Hannes Westermann, and Leila Kosseim. 2021. Extracting facts from case rulings through paragraph segmentation of judicial decisions. In *International Conference on Applications of Natural Language to Information Systems*, pages 187–198. Springer.

Yixiao Ma, Yueyue Wu, Qingyao Ai, Yiqun Liu, Yunqiu Shao, Min Zhang, and Shaoping Ma. 2024. Incorporating structural information into legal case retrieval. *ACM Transactions on Information Systems*, 42(2):40:1–40:28.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22:216–242.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Kumar Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shubham Kumar Nigam, Tanmay Dubey, Govind Sharma, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1129–1144, Albuquerque, New Mexico. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

TYS S Santosh, Philipp Bock, and Matthias Grabmair. 2023. Joint span segmentation and rhetorical role labeling with data augmentation for legal documents. In *European Conference on Information Retrieval*, pages 627–636. Springer.

TYSS Santosh, Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306.

Jaromir Savelka and Kevin D Ashley. 2018. Segmenting U.S. court decisions into functional and issue specific parts. In *Legal Knowledge and Information Systems: JURIX 2018*, pages 111–120. IOS Press.

Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. volume 25, pages 107–126.

Kushagri Tandon and Niladri Chatterjee. 2023. LRL_NC at SemEval-2023 task 6: Sequential sentence classification for legal documents using topic modeling features. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 143–149, Toronto, Canada. Association for Computational Linguistics.

Julien Van den Avenne. 2024. Reynaerde-7B-Chat.

Romy AN van Drie, Maaike HT de Boer, Roos M Bakker, Ioannis Tolios, and Daan Vos. 2023. The dutch law as a semantic role labeling dataset. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 316–322.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. *ASAIL@ ICAIL*, 2385:1–10.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *Legal Knowledge and Information Systems: JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8–10 December 2021*, pages 33–42. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP.

Yang Zhong and Diane Litman. 2022. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 46–51. Association for Computational Linguistics.

Appendix

A Annotation Scheme

Guidelines – in Dutch

Annoteer een gerechtelijke uitspraak, of een onderdeel ervan, enkel indien deze gaat over de (1) materële feiten, (2) beoordeling en (3) beslissing. Deze (en andere) categorieën kunnen als volgt worden omschreven:

- **Materële feiten.** Feiten die rechtstreeks verband houden met de inhoudelijke beoordeling. Ze vormen de feitelijke grondslag van een vordering of verweer, zijn relevant voor of een regel van toepassing is bepalend voor de beoordeling en beslissing van de rechter.
- **Beoordeling.** De juridische analyse en motivering waarin de rechter uitlegt hoe en waarom deze tot een bepaalde beslissing komt.
- **Beslissing.** Geeft de beslissing of uitspraak van de rechter.
- **Procesverloop/proceshandelingen.** Processtappen en formele handelingen, inclusief de standpunten van partijen, wat zij hebben gevorderd, verzocht, verweerd, tenlastegelegd, onderbouwd, of wat zij aan stukken hebben ingebracht, die niet inhoudelijk ingrijpen op de beoordeling van het geschil.

- **Beslissingen in eerdere instanties.** Wat rechters in eerdere instanties van de zaak hebben beslist.
- **Wettelijk kader.** Hieronder vallen de toepasselijke wetsbepalingen, relevante jurisprudentie en andere toepasselijke regels.
- **Proceskosten.** Het gaat hierbij over de beslissing wie welke proceskosten dient te dragen.
- **Overig.** Informatie die niet onder een van bovenstaande categorieën kan worden gebracht.

Afhankelijk van de aard van de zaak, kan informatie die normaal gesproken tot het procesverloop of tot proceshandelingen behoort, toch als materieel feit worden aangemerkt. Bijvoorbeeld:

- *"De dagvaarding is niet ontvangen."*
  Normaal gesproken gaat dit over een vormfout in de procedure (=procesverloop). In een klacht over schending van het recht op verdediging, kan het een materieel feit zijn (want de procedurele stappen bepalen de inhoudelijke beoordeling).
- *"Op 19 mei 2009 heeft Het Regionaal Tuchtcollege de maatregel van een waarschuwing opgelegd."*
  Dit kan een procedureel feit zijn, maar in een zaak over de aansprakelijkheid van een ziekenhuis/arts is dit feit van belang voor de beoordeling van het geschil.
- *"Het Hof Amsterdam heeft gedaagde in de strafzaak veroordeeld."*
  Dit kan een procedureel feit zijn (bv. in cassatie), maar als het gaat om de aansprakelijkheid van de verdachte, is het een materieel feit (want de veroordeling is relevant voor de aansprakelijkheid).

Het annoteren dient te gebeuren door een of meerdere zinnen te annoteren.

Excluded Categories

The full annotation scheme distinguishes eight categories. The following four were excluded as they fall outside the core argumentative and decisional structure of a judicial decision. Decisions in lower courts (*beslissingen in eerdere instanties*) provide background context rather than forming part of the current court's own reasoning. The legal framework (*wettelijk kader*) — comprising applicable statutory provisions, relevant case law, and other applicable rules —serves as reference material supporting the court's reasoning rather than constituting reasoning itself. Legal costs(*proceskosten*) concern the administrative decision as to who bears which costs rather than the substantive reasoning or outcome of the case. The residual category (*overig*) covers information that cannot be classified under any of the above categories and is therefore not part of the legal substance of the decision.

Preprocessing and Sentence Segmentation

Before segmentation, the text was conservatively normalized while preserving relevant newline and layout cues. This included the removal of control characters and formatting artefacts, the repair of line-break hyphenation (e.g., woord-vervolg → woordvervolg), and the normalization of irregular spacing (e.g., woord , → woord,). Sentences were identified using punctuation-based rules combined with abbreviation guards and layout-sensitive heuristics. Specifically: abbreviations such as art., mr., and bijv. were not treated as sentence boundaries; headings such as FEITEN or BEOORDELING were preserved as separate units; list prefixes such as 1. or (a) were treated as structural markers rather than ordinary sentence content and were removed; and punctuation inside brackets was protected during splitting so that references such as (art. 6:162 BW) would not trigger incorrect sentence breaks. After segmentation, numbering and bullet points were removed and whitespace was normalized to produce the final sentence-level input for classification. Sentences resulting from clear segmentation errors, identified during header mapping, were removed prior to experimentation. In total, 70 sentences were excluded, representing less than 1% of the original total. The resulting final dataset contained 8,391 sentences.

Header Mapping

Raw section titles were normalized to broader functional groups using a manually constructed seed word mapping. Each raw header was matched against the seed words list and assigned to the corresponding group. The mapping covers five retained groups and three excluded categories, listed below.

Retained header groups:

- **Context:** context_overig, voorvragen, de

voorvragen, inleiding, grondslag en inhoud van het eab, onderzoek van de zaak, het onderzoek ter terechtzitting, onderzoek ter terechtzitting, gronden, het onderzoek op de terechtzitting, onderzoek op de terechtzitting, de kern van de zaak, waar gaat deze zaak over?, inleiding en samenvatting, waar gaat het over?, het bewijs, waar gaat de zaak over?

- **Feiten:** feiten, de feiten, de zaak in het kort, de vaststaande feiten, vaststaande feiten, uitgangspunten en feiten, feiten en procesverloop, feitelijke achtergrond, uitgangspunten in cassatie

- **Beoordeling:** beoordeling_door_rechter, de beoordeling, waardering van het bewijs, de motivering van de beslissing, beoordeling, bewezenverklaring, de beoordeling van het bewijs, de verdere beoordeling, beoordeling van het geschil, de strafbaarheid van de verdachte, strafbaarheid van verdachte, de strafbaarheid van verdachte, de strafbaarheid, beoordeling door de rechtbank, strafbaarheid, beoordeling van het middel, strafbaarheid van de feiten, de strafbaarheid van het bewezenverklaarde, conclusie en gevolgen, de bewezenverklaring, de strafbaarheid van de feiten, de strafbaarheid van het bewezen verklaarde, beoordeling van het cassatiemiddel, beoordeling van de middelen, bespreking van het cassatiemiddel, de bewijsmotivering, beoordeling van het bewijs, strafbaarheid verdachte, de motivering van de beslissing in hoger beroep, beoordeling van de ontvankelijkheid van het beroep in cassatie, de bewijsbeslissing, de kwalificatie van het bewezenverklaarde, strafbaarheid van de verdachte, de gronden van de beslissing, motivering van de straf, de strafbaarheid van het feit, de beoordeling van het geschil, strafbaarheid van het feit, motivering straf, motivering van de sanctie, het oordeel van de rechtbank, de beoordeling in het incident, beoordeling van de klachten, de beoordeling in hoger beroep, overwegingen, kwalificatie en strafbaarheid van de feiten, beoordeling van de cassatiemiddelen, beoordeling van het eerste cassatiemiddel, strafbaarheid feiten, de overwegingen ten aanzien van straf en/of maatregel, beoordeling van het tweede cassatiemiddel, beoordeling in hoger beroep, strafbaarheid feit, motivering van de straffen en maatregelen, kwalificatie en strafbaarheid van het feit, de beoordeling in conventie en in reconventie, het geschil in reconventie, de straf en/of de maatregel, bewijs, het eerste middel, de beoordeling van de civiele vordering

- **Beslissing:** beslissing, de beslissing, slotsom, de uitspraak, de slotsom, de strafoplegging, conclusie, oplegging van straf, de op te leggen straf of maatregel, vrijspraak, .beslissing, . beslissing, de straf

- **Proceshandelingen partijen:** proceshandelingen_partijen, het geschil, tenlastelegging, de tenlastelegging, de vordering, het verweer, het verzoek, geschil, geding in cassatie, eis officier van justitie, de vordering en het verweer, de inhoud van de tenlastelegging, het wrakingsverzoek, geschil in hoger beroep, het verzoek en het verweer, geschil en conclusies van partijen, het cassatieberoep, het geschil in conventie, de standpunten, standpunten van partijen, het standpunt van de officier van justitie, de vordering tot tenuitvoerlegging

Excluded categories:

- **Wettelijk kader:** wettelijk_kader, toepasselijke wettelijke voorschriften, de wettelijke voorschriften, toepassende wetsbepalingen, toepasselijke wetsartikelen, de toegepaste wettelijke voorschriften, de toepasselijke wetsartikelen, de toegepaste wettelijke bepalingen

- **Proceskosten:** proceskosten, griffierecht en proceskosten, kosten, proceskosten en griffierecht

- **Procesverloop:** procesverloop, de procedure, het geding in hoger beroep, het procesverloop, het geding in eerste aanleg, ontstaan en loop van het geding, procesgang, procesverloop in cassatie, het verloop van de procedure, het verloop van het geding in eerste aanleg, het verdere verloop van het geding in hoger beroep, het verloop van het geding, het verloop van de procedure in hoger beroep, de procedure in hoger beroep, procesverloop in hoger beroep, procedure, het verdere verloop van de procedure in hoger beroep, het verdere verloop van de procedure, de procedure bij de rechtbank, de procedure in eerste aanleg, verloop van de procedure, het verdere procesverloop, procedure bij de rechtbank, het geschil en de beslissing in eerste aanleg

B LLM Prompt Templates

This appendix reports the prompt templates used for the instruction-tuned LLM experiments. The same Dutch label instruction block was used in all prompting conditions. The placeholders <SENTENCE>, <FEW_SHOT_BLOCK>, <HDR_TITLE>, and <HDR_GROUP> were replaced at runtime with the corresponding sentence, few-shot examples, raw section header, and normalized header group.

B.1 Shared label instruction block

```
Je bent een annotator van Nederlandse
↪  rechterlijke uitspraken.

Je taak is om één zin uit een rechterlijke
↪  uitspraak te classificeren onder precies één
↪  van de volgende labels:

"None", "materiele feiten", "proceshandelingen",
↪  "beoordeling", "beslissing".

Kies altijd precies één label en geef uitsluitend
↪  het label als antwoord.

Labeldefinities:

1. "materiele feiten"
Gebruik dit label voor feiten die rechtstreeks
↪  verband houden met de inhoudelijke
↪  beoordeling van het geschil. Het gaat om
↪  feiten die de feitelijke grondslag vormen van
↪  een vordering of verweer, relevant zijn voor
↪  de toepasselijkheid van een rechtsregel, of
↪  bepalend zijn voor de beoordeling en
↪  beslissing van de rechter.

Let op: informatie die op het eerste gezicht
↪  procedureel lijkt, kan toch "materiele
↪  feiten" zijn als die procedurele gebeurtenis
↪  zelf inhoudelijk relevant is voor het
↪  geschil. Bijvoorbeeld: het niet ontvangen
↪  van een dagvaarding kan een materieel feit
↪  zijn als de zaak gaat over schending van het
↪  recht op verdediging.

2. "beoordeling"
Gebruik dit label voor de juridische analyse en
↪  motivering van de rechter. Dit omvat de
↪  interpretatie van feiten of regels,
↪  bewijswaardering, toepassing van het recht
↪  op de feiten, en uitleg waarom de rechter tot
↪  een bepaalde beslissing komt.

3. "beslissing"
Gebruik dit label voor de beslissing of uitspraak
↪  van de rechter zelf. Dit omvat het dictum,
↪  toewijzing of afwijzing van vorderingen of
↪  verzoeken, gegrond- of ongegrondverklaringen,
↪  veroordelingen, vrijspraken, opgelegde
↪  maatregelen en andere bindende beslissingen
↪  van de rechter.

4. "proceshandelingen"
Gebruik dit label voor standpunten en formele
↪  proceshandelingen van partijen binnen de
↪  procedure, zoals wat partijen hebben
↪  gevorderd, verzocht, aangevoerd, betwist,
↪  verweerd, ten laste gelegd, onderbouwd, of
↪  aan stukken hebben ingebracht, voor zover
↪  deze informatie niet rechtstreeks
↪  inhoudelijk bepalend is voor de beoordeling
↪  van het geschil.

Belangrijk: algemeen procesverloop valt niet
↪  automatisch onder "proceshandelingen".
↪  Zinnen die alleen beschrijven hoe de
↪  procedure is verlopen, zoals het instellen
↪  van hoger beroep, het houden van een zitting,
↪  het indienen of ontvangen van stukken,
↪  verwijzingen naar eerdere processtappen, of
↪  het verloop bij eerdere instanties, krijgen
↪  in principe "None", tenzij de procedurele
↪  gebeurtenis zelf inhoudelijk relevant is voor
↪  het geschil. In dat geval gebruik je
↪  "materiele feiten".

5. "None"
Gebruik dit label voor zinnen die niet onder de
↪  vier bovenstaande inhoudelijke categorieën
↪  vallen. Dit omvat onder meer algemene
↪  inleidingen, metadata, kopjes, verwijzingen,
↪  algemeen procesverloop, beslissingen in
↪  eerdere instanties, wettelijk kader,
↪  proceskosten en overige informatie die niet
↪  direct als "materiele feiten",
↪  "proceshandelingen", "beoordeling" of
↪  "beslissing" moet worden geclassificeerd.

Beslisregels bij twijfel:

- Gaat de zin over de inhoudelijke feiten van het
↪  geschil? Kies "materiele feiten".
- Legt de rechter juridisch uit waarom iets wel
↪  of niet geldt? Kies "beoordeling".
- Spreekt de rechter een bindende uitkomst uit?
↪  Kies "beslissing".
- Beschrijft de zin wat een partij vordert,
↪  stelt, betwist, verzoekt of heeft ingebracht,
↪  zonder dat dit zelf als inhoudelijk feit
↪  wordt gebruikt? Kies "proceshandelingen".
- Beschrijft de zin alleen algemeen
↪  procesverloop, wettelijk kader, proceskosten,
↪  eerdere beslissingen of overige context?
↪  Kies "None".
```

B.2 Zero-shot prompt

```
Je bent een annotator van Nederlandse
↪  rechterlijke uitspraken.

Je taak is om één zin uit een rechterlijke
↪  uitspraak te classificeren onder precies één
↪  van de volgende labels:

"None", "materiele feiten", "proceshandelingen",
↪  "beoordeling", "beslissing".

Kies altijd precies één label en geef uitsluitend
↪  het label als antwoord.

Labeldefinities:

1. "materiele feiten"
```

```
Gebruik dit label voor feiten die rechtstreeks
↪ verband houden met de inhoudelijke
↪ beoordeling van het geschil. Het gaat om
↪ feiten die de feitelijke grondslag vormen van
↪ een vordering of verweer, relevant zijn voor
↪ de toepasselijkheid van een rechtsregel, of
↪ bepalend zijn voor de beoordeling en
↪ beslissing van de rechter.

Let op: informatie die op het eerste gezicht
↪ procedureel lijkt, kan toch "materiele
↪ feiten" zijn als die procedurele gebeurtenis
↪ zelf inhoudelijk relevant is voor het
↪ geschil. Bijvoorbeeld: het niet ontvangen
↪ van een dagvaarding kan een materieel feit
↪ zijn als de zaak gaat over schending van het
↪ recht op verdediging.

2. "beoordeling"
Gebruik dit label voor de juridische analyse en
↪ motivering van de rechter. Dit omvat de
↪ interpretatie van feiten of regels,
↪ bewijswaardering, toepassing van het recht
↪ op de feiten, en uitleg waarom de rechter tot
↪ een bepaalde beslissing komt.

3. "beslissing"
Gebruik dit label voor de beslissing of uitspraak
↪ van de rechter zelf. Dit omvat het dictum,
↪ toewijzing of afwijzing van vorderingen of
↪ verzoeken, gegrond- of ongegrondverklaringen,
↪ veroordelingen, vrijspraken, opgelegde
↪ maatregelen en andere bindende beslissingen
↪ van de rechter.

4. "proceshandelingen"
Gebruik dit label voor standpunten en formele
↪ proceshandelingen van partijen binnen de
↪ procedure, zoals wat partijen hebben
↪ gevorderd, verzocht, aangevoerd, betwist,
↪ verweerd, ten laste gelegd, onderbouwd, of
↪ aan stukken hebben ingebracht, voor zover
↪ deze informatie niet rechtstreeks
↪ inhoudelijk bepalend is voor de beoordeling
↪ van het geschil.

Belangrijk: algemeen procesverloop valt niet
↪ automatisch onder "proceshandelingen".
↪ Zinnen die alleen beschrijven hoe de
↪ procedure is verlopen, zoals het instellen
↪ van hoger beroep, het houden van een zitting,
↪ het indienen of ontvangen van stukken,
↪ verwijzingen naar eerdere processtappen, of
↪ het verloop bij eerdere instanties, krijgen
↪ in principe "None", tenzij de procedurele
↪ gebeurtenis zelf inhoudelijk relevant is voor
↪ het geschil. In dat geval gebruik je
↪ "materiele feiten".

5. "None"
Gebruik dit label voor zinnen die niet onder de
↪ vier bovenstaande inhoudelijke categorieën
↪ vallen. Dit omvat onder meer algemene
↪ inleidingen, metadata, kopjes, verwijzingen,
↪ algemeen procesverloop, beslissingen in
↪ eerdere instanties, wettelijk kader,
↪ proceskosten en overige informatie die niet
↪ direct als "materiele feiten",
↪ "proceshandelingen", "beoordeling" of
↪ "beslissing" moet worden geclassificeerd.

Beslisregels bij twijfel:

- Gaat de zin over de inhoudelijke feiten van het
↪ geschil? Kies "materiele feiten".
- Legt de rechter juridisch uit waarom iets wel
↪ of niet geldt? Kies "beoordeling".
- Spreekt de rechter een bindende uitkomst uit?
↪ Kies "beslissing".
- Beschrijft de zin wat een partij vordert,
↪ stelt, betwist, verzoekt of heeft ingebracht,
↪ zonder dat dit zelf als inhoudelijk feit
↪ wordt gebruikt? Kies "proceshandelingen".
- Beschrijft de zin alleen algemeen
↪ procesverloop, wettelijk kader, proceskosten,
↪ eerdere beslissingen of overige context?
↪ Kies "None".

Classificeer de volgende zin.

Zin:
"""<SENTENCE>"""

Geef alleen geldige JSON terug in exact dit
↪ formaat:
{
  "label": "None | materiele feiten |
  ↪ proceshandelingen | beoordeling |
  ↪ beslissing"
}
```

B.3 Few-shot block format

```
Voorbeelden:

Zin:
"""<FEW-SHOT SENTENCE>"""

Antwoord:
{"label": "<FEW-SHOT LABEL>"}
```

B.4 Few-shot prompt

```
Je bent een annotator van Nederlandse
↪ rechterlijke uitspraken.

Je taak is om één zin uit een rechterlijke
↪ uitspraak te classificeren onder precies één
↪ van de volgende labels:

"None", "materiele feiten", "proceshandelingen",
↪ "beoordeling", "beslissing".

Kies altijd precies één label en geef uitsluitend
↪ het label als antwoord.

Labeldefinities:

1. "materiele feiten"
Gebruik dit label voor feiten die rechtstreeks
↪ verband houden met de inhoudelijke
↪ beoordeling van het geschil. Het gaat om
↪ feiten die de feitelijke grondslag vormen van
↪ een vordering of verweer, relevant zijn voor
↪ de toepasselijkheid van een rechtsregel, of
↪ bepalend zijn voor de beoordeling en
↪ beslissing van de rechter.
```

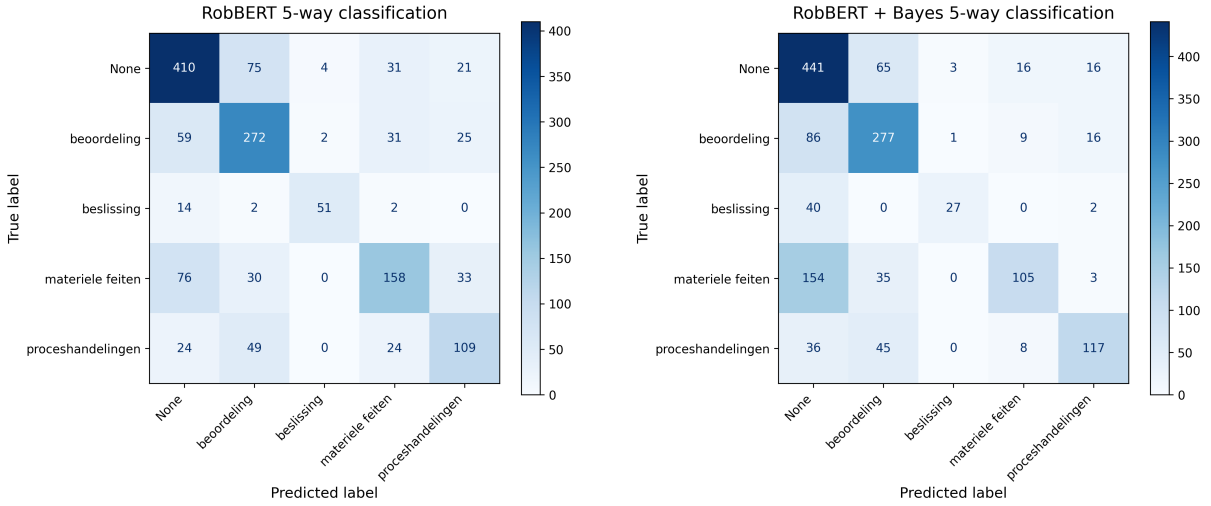
```
Let op: informatie die op het eerste gezicht
↪  procedureel lijkt, kan toch "materiele
↪  feiten" zijn als die procedurele gebeurtenis
↪  zelf inhoudelijk relevant is voor het
↪  geschil. Bijvoorbeeld: het niet ontvangen
↪  van een dagvaarding kan een materieel feit
↪  zijn als de zaak gaat over schending van het
↪  recht op verdediging.

2. "beoordeling"
Gebruik dit label voor de juridische analyse en
↪  motivering van de rechter. Dit omvat de
↪  interpretatie van feiten of regels,
↪  bewijswaardering, toepassing van het recht
↪  op de feiten, en uitleg waarom de rechter tot
↪  een bepaalde beslissing komt.

3. "beslissing"
Gebruik dit label voor de beslissing of uitspraak
↪  van de rechter zelf. Dit omvat het dictum,
↪  toewijzing of afwijzing van vorderingen of
↪  verzoeken, gegrond- of ongegrondverklaringen,
↪  veroordelingen, vrijspraken, opgelegde
↪  maatregelen en andere bindende beslissingen
↪  van de rechter.

4. "proceshandelingen"
Gebruik dit label voor standpunten en formele
↪  proceshandelingen van partijen binnen de
↪  procedure, zoals wat partijen hebben
↪  gevorderd, verzocht, aangevoerd, betwist,
↪  verweerd, ten laste gelegd, onderbouwd, of
↪  aan stukken hebben ingebracht, voor zover
↪  deze informatie niet rechtstreeks
↪  inhoudelijk bepalend is voor de beoordeling
↪  van het geschil.

Belangrijk: algemeen procesverloop valt niet
↪  automatisch onder "proceshandelingen".
↪  Zinnen die alleen beschrijven hoe de
↪  procedure is verlopen, zoals het instellen
↪  van hoger beroep, het houden van een zitting,
↪  het indienen of ontvangen van stukken,
↪  verwijzingen naar eerdere processtappen, of
↪  het verloop bij eerdere instanties, krijgen
↪  in principe "None", tenzij de procedurele
↪  gebeurtenis zelf inhoudelijk relevant is voor
↪  het geschil. In dat geval gebruik je
↪  "materiele feiten".

5. "None"
Gebruik dit label voor zinnen die niet onder de
↪  vier bovenstaande inhoudelijke categorieën
↪  vallen. Dit omvat onder meer algemene
↪  inleidingen, metadata, kopjes, verwijzingen,
↪  algemeen procesverloop, beslissingen in
↪  eerdere instanties, wettelijk kader,
↪  proceskosten en overige informatie die niet
↪  direct als "materiele feiten",
↪  "proceshandelingen", "beoordeling" of
↪  "beslissing" moet worden geclassificeerd.

Beslisregels bij twijfel:

- Gaat de zin over de inhoudelijke feiten van het
↪  geschil? Kies "materiele feiten".
- Legt de rechter juridisch uit waarom iets wel
↪  of niet geldt? Kies "beoordeling".
- Spreekt de rechter een bindende uitkomst uit?
↪  Kies "beslissing".
- Beschrijft de zin wat een partij vordert,
↪  stelt, betwist, verzoekt of heeft ingebracht,
↪  zonder dat dit zelf als inhoudelijk feit
↪  wordt gebruikt? Kies "proceshandelingen".
- Beschrijft de zin alleen algemeen
↪  procesverloop, wettelijk kader, proceskosten,
↪  eerdere beslissingen of overige context?
↪  Kies "None".

<FEW_SHOT_BLOCK>

Classificeer nu de volgende zin.

Zin:
"""<SENTENCE>"""

Geef alleen geldige JSON terug in exact dit
↪  formaat:
{
  "label": "None | materiele feiten |
  ↪  proceshandelingen | beoordeling |
  ↪  beslissing"
}
```

B.5 Header-aware few-shot prompt

```
Je bent een annotator van Nederlandse
↪  rechterlijke uitspraken.

Je taak is om één zin uit een rechterlijke
↪  uitspraak te classificeren onder precies één
↪  van de volgende labels:

"None", "materiele feiten", "proceshandelingen",
↪  "beoordeling", "beslissing".

Kies altijd precies één label en geef uitsluitend
↪  het label als antwoord.

Labeldefinities:

1. "materiele feiten"
Gebruik dit label voor feiten die rechtstreeks
↪  verband houden met de inhoudelijke
↪  beoordeling van het geschil. Het gaat om
↪  feiten die de feitelijke grondslag vormen van
↪  een vordering of verweer, relevant zijn voor
↪  de toepasselijkheid van een rechtsregel, of
↪  bepalend zijn voor de beoordeling en
↪  beslissing van de rechter.

Let op: informatie die op het eerste gezicht
↪  procedureel lijkt, kan toch "materiele
↪  feiten" zijn als die procedurele gebeurtenis
↪  zelf inhoudelijk relevant is voor het
↪  geschil. Bijvoorbeeld: het niet ontvangen
↪  van een dagvaarding kan een materieel feit
↪  zijn als de zaak gaat over schending van het
↪  recht op verdediging.

2. "beoordeling"
Gebruik dit label voor de juridische analyse en
↪  motivering van de rechter. Dit omvat de
↪  interpretatie van feiten of regels,
↪  bewijswaardering, toepassing van het recht
↪  op de feiten, en uitleg waarom de rechter tot
↪  een bepaalde beslissing komt.

3. "beslissing"
```

```
Gebruik dit label voor de beslissing of uitspraak
↪  van de rechter zelf. Dit omvat het dictum,
↪  toewijzing of afwijzing van vorderingen of
↪  verzoeken, gegrond- of ongegrondverklaringen,
↪  veroordelingen, vrijspraken, opgelegde
↪  maatregelen en andere bindende beslissingen
↪  van de rechter.

4. "proceshandelingen"
Gebruik dit label voor standpunten en formele
↪  proceshandelingen van partijen binnen de
↪  procedure, zoals wat partijen hebben
↪  gevorderd, verzocht, aangevoerd, betwist,
↪  verweerd, ten laste gelegd, onderbouwd, of
↪  aan stukken hebben ingebracht, voor zover
↪  deze informatie niet rechtstreeks
↪  inhoudelijk bepalend is voor de beoordeling
↪  van het geschil.

Belangrijk: algemeen procesverloop valt niet
↪  automatisch onder "proceshandelingen".
↪  Zinnen die alleen beschrijven hoe de
↪  procedure is verlopen, zoals het instellen
↪  van hoger beroep, het houden van een zitting,
↪  het indienen of ontvangen van stukken,
↪  verwijzingen naar eerdere processtappen, of
↪  het verloop bij eerdere instanties, krijgen
↪  in principe "None", tenzij de procedurele
↪  gebeurtenis zelf inhoudelijk relevant is voor
↪  het geschil. In dat geval gebruik je
↪  "materiele feiten".

5. "None"
Gebruik dit label voor zinnen die niet onder de
↪  vier bovenstaande inhoudelijke categorieën
↪  vallen. Dit omvat onder meer algemene
↪  inleidingen, metadata, kopjes, verwijzingen,
↪  algemeen procesverloop, beslissingen in
↪  eerdere instanties, wettelijk kader,
↪  proceskosten en overige informatie die niet
↪  direct als "materiele feiten",
↪  "proceshandelingen", "beoordeling" of
↪  "beslissing" moet worden geclassificeerd.

Beslisregels bij twijfel:

- Gaat de zin over de inhoudelijke feiten van het
↪  geschil? Kies "materiele feiten".
- Legt de rechter juridisch uit waarom iets wel
↪  of niet geldt? Kies "beoordeling".
- Spreekt de rechter een bindende uitkomst uit?
↪  Kies "beslissing".
- Beschrijft de zin wat een partij vordert,
↪  stelt, betwist, verzoekt of heeft ingebracht,
↪  zonder dat dit zelf als inhoudelijk feit
↪  wordt gebruikt? Kies "proceshandelingen".
- Beschrijft de zin alleen algemeen
↪  procesverloop, wettelijk kader, proceskosten,
↪  eerdere beslissingen of overige context?
↪  Kies "None".

<FEW_SHOT_BLOCK>

Je krijgt nu een zin uit een Nederlandse
↪  rechterlijke uitspraak.
De sectiekop kan context geven, maar de zin zelf
↪  blijft het belangrijkste bewijs.
Baseer het label dus niet alleen op de sectiekop.

Originele sectiekop:
"""<HDR_TITLE>"""

Genormaliseerde kopgroep:
"""<HDR_GROUP>"""

Zin:
"""<SENTENCE>"""

Geef alleen geldige JSON terug in exact dit
↪  formaat:
{
  "label": "None | materiele feiten |
  ↪  proceshandelingen | beoordeling |
  ↪  beslissing"
}
```

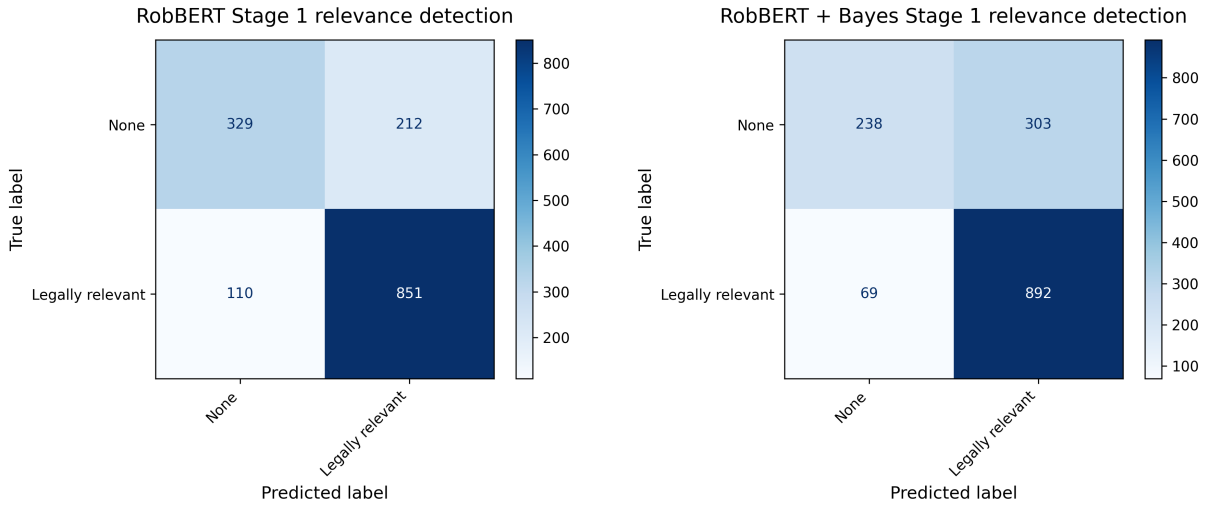
C Additional Results



(a) RobBERT

(b) RobBERT + Bayes

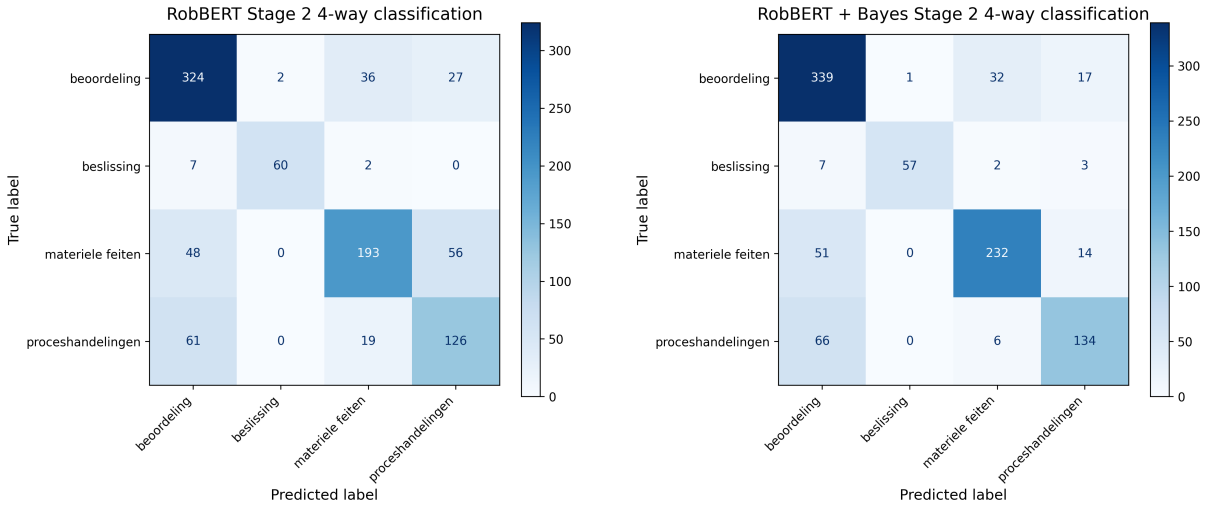
Figure 3: Confusion matrices for direct five-way classification. Rows indicate true labels and columns indicate predicted labels.



(a) RobBERT

(b) RobBERT + Bayes

Figure 4: Confusion matrices for Stage 1 relevance detection. Rows indicate true labels and columns indicate predicted labels.



(a) RobBERT

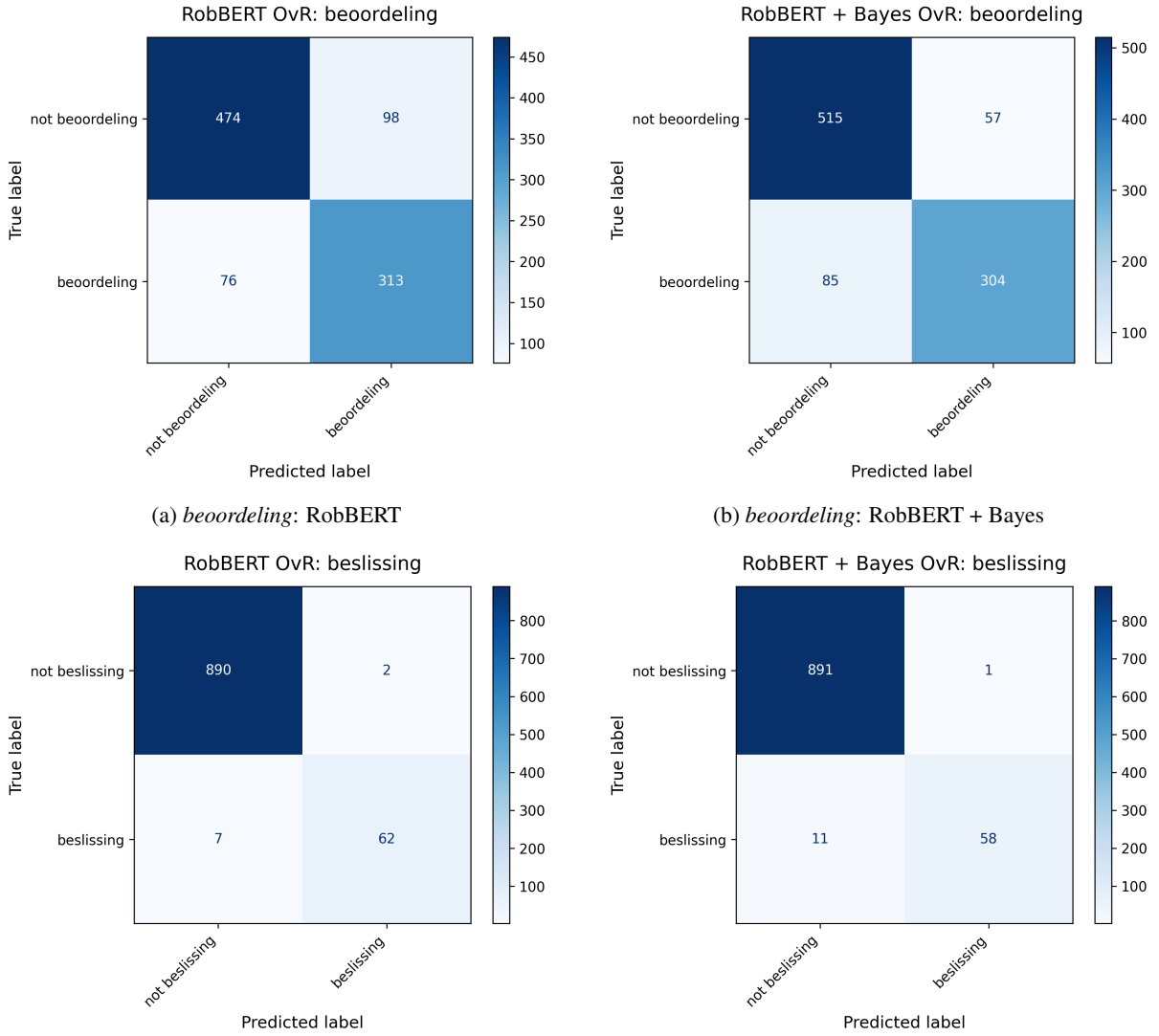
(b) RobBERT + Bayes

Figure 5: Confusion matrices for Stage 2 four-way role classification. Rows indicate true labels and columns indicate predicted labels.

Table 5: Overall supervised model performance before and after Bayesian updating

Model	Setting	Evaluation subset	Support	Before Bayes					After Bayes				
				Acc.	Macro P	Macro R	Macro F1	Weighted F1	Acc.	Macro P	Macro R	Macro F1	Weighted F1
Logistic Regression	5-way	All sentences	1502	0.5573	0.5777	0.5716	0.5726	0.5555	0.5033	0.6586	0.4462	0.4439	0.4578
	Stage 1	All sentences, binary	1502	0.6851	0.6686	0.6792	0.6704	0.6899	0.6711	0.6349	0.6093	0.6116	0.6541
	Stage 2 4-way	Legal-labelled only	961	0.6795	0.7094	0.6905	0.6970	0.6754	0.7086	0.7989	0.7255	0.7429	0.7056
	OvR mean	Legal-labelled only	961	0.8400	0.7934	0.7834	0.7866	0.8384	0.8736	0.8801	0.7795	0.8116	0.8657
	Pipeline A	All sentences	1502	0.5233	0.6033	0.4703	0.4949	0.4921	0.4880	0.5637	0.5212	0.4926	0.4496
	Pipeline B	All sentences	1502	0.5120	0.6035	0.4400	0.4665	0.4752	0.4880	0.5776	0.5179	0.4953	0.4498
BERTje	5-way	All sentences	1502	0.6458	0.6774	0.6398	0.6561	0.6442	0.6525	0.7407	0.6166	0.6549	0.6469
	Stage 1	All sentences, binary	1502	0.7783	0.7595	0.7589	**0.7592**	0.7782	0.7790	0.7616	0.7513	0.7556	0.7768
	Stage 2 4-way	Legal-labelled only	961	0.7138	0.7484	0.7128	0.7272	0.7130	0.7846	0.8277	0.7741	0.7965	0.7838
	OvR mean	Legal-labelled only	961	0.8606	0.8336	0.7965	0.8118	0.8579	0.8840	0.8886	0.7974	0.8270	0.8764
	Pipeline A	All sentences	1502	0.6258	0.6466	0.5997	0.6164	0.6235	0.6605	0.7036	0.6485	**0.6679**	0.6595
	Pipeline B	All sentences	1502	0.6272	0.6530	0.6028	0.6212	0.6256	0.6678	0.7110	0.6513	**0.6717**	0.6672
RobBERT	5-way	All sentences	1502	0.6658	0.6911	0.6515	**0.6680**	0.6634	0.6438	0.7261	0.5680	0.6071	0.6311
	Stage 1	All sentences, binary	1502	0.7856	0.7750	0.7468	0.7562	0.7799	0.7523	0.7608	0.6841	0.6944	0.7316
	Stage 2 4-way	Legal-labelled only	961	0.7315	0.7697	0.7410	0.7526	0.7304	0.7929	0.8414	0.7823	**0.8064**	0.7922
	OvR mean	Legal-labelled only	961	0.8733	0.8472	0.8224	0.8328	0.8709	0.8881	0.9020	0.8081	**0.8351**	0.8794
	Pipeline A	All sentences	1502	0.6372	0.6615	0.6498	0.6513	0.6380	0.6252	0.6795	0.6629	0.6539	0.6201
	Pipeline B	All sentences	1502	0.6372	0.6705	0.6567	0.6604	0.6377	0.6352	0.6920	0.6847	0.6711	0.6289

Note. Stage 1 is evaluated as a binary classification task distinguishing unlabelled sentences from legally labelled sentences. Stage 2 is evaluated only on gold legal-labelled sentences. OvR mean refers to the arithmetic average over the four binary one-vs-rest classifiers: *beoordeling*, *beslissing*, *materiële feiten*, and *proceshandelingen*. Pipeline A combines Stage 1 and Stage 2 four-way probabilities; Pipeline B combines Stage 1 and normalized OvR probabilities. For pipeline variants, After Bayes applies Bayesian updating to both components independently before combination. Bayesian updating combines model probabilities with header-based priors estimated from the training set. Best Macro F1 scores within each setting are shown in bold.



(a) *beoordeling*: RobBERT

(b) *beoordeling*: RobBERT + Bayes

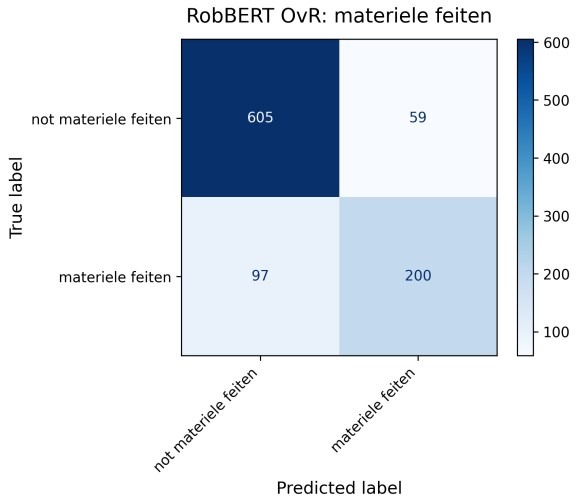
(c) *beslissing*: RobBERT

(d) *beslissing*: RobBERT + Bayes

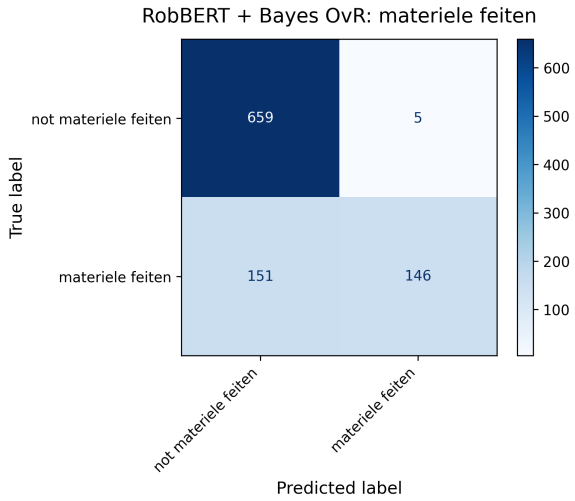
Figure 6: Confusion matrices for OvR classifiers for *beoordeling* and *beslissing*. Rows indicate true labels and columns indicate predicted labels.

Table 6: Prompting-based LLM performance on test set. The prompting experiments are evaluated directly on the frozen test set (1502 sentences) without task-specific fine-tuning. Bayesian updating is not applied to these models because the prompting setup produces generated label responses rather than calibrated class probabilities.

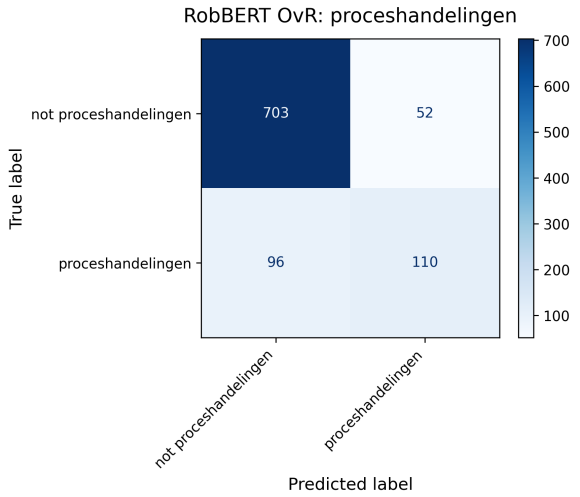
Model	**Setting**	**Acc.**	**Macro P**	**Macro R**	**Macro F1**	**Weighted F1**
Gemma-3-12B-IT	Zero-shot	0.4970	0.4835	0.5483	0.4938	0.4964
Gemma-3-12B-IT	Few-shot	0.4827	0.4641	0.4951	0.4707	0.4781
Gemma-3-12B-IT	Header-aware few-shot	0.4933	0.4850	0.5159	0.4935	0.4892
Qwen2.5-7B-Instruct	Zero-shot	0.3216	0.3666	0.3771	0.3268	0.3245
Qwen2.5-7B-Instruct	Few-shot	0.4021	0.3942	0.4391	0.4069	0.4046
Qwen2.5-7B-Instruct	Header-aware few-shot	0.4221	0.4160	0.4754	0.4288	0.4061
Reynaerde-7B-Chat	Zero-shot	0.3655	0.3703	0.3299	0.2888	0.3102
Reynaerde-7B-Chat	Few-shot	0.3628	0.3570	0.3042	0.3126	0.3435
Reynaerde-7B-Chat	Header-aware few-shot	0.4081	0.4405	0.4558	0.4275	0.4037



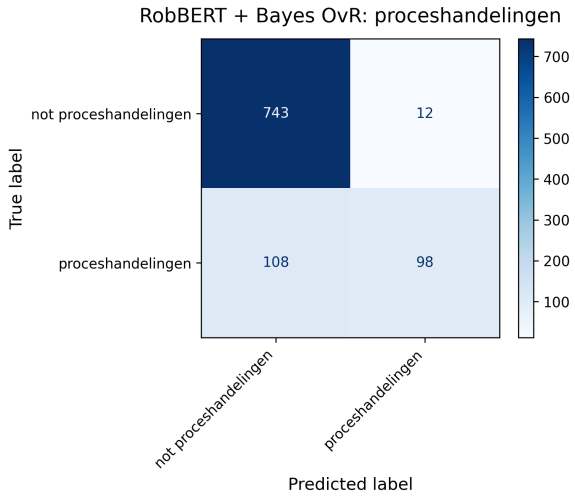
(a) *materiële feiten*: RobBERT



(b) *materiële feiten*: RobBERT + Bayes



(c) *proceshandelingen*: RobBERT



(d) *proceshandelingen*: RobBERT + Bayes

Figure 7: Confusion matrices for OvR classifiers for *materiële feiten* and *proceshandelingen*. Rows indicate true labels and columns indicate predicted labels.