TraffickingLens: Evaluating LLMs for Legal Element Extraction from Human Trafficking Case Narratives

Xihao Cao
Boston University
Boston, MA, USA
xihaoc@bu.edu

Abstract

Anti-human-trafficking organizations and investigators routinely work with large amounts of narrative text, but converting them into structured information remains costly and difficult at scale. We introduce **TraffickingLens**, a benchmark and evaluation framework for studying how effectively large language models (LLMs) can structure human-trafficking narratives. Our evaluation uses 1,263 English case narratives from the United Nations Office on Drugs and Crime (UNODC) SHERLOC database. We compare a sparse supervised classifier and a fine-tuned Transformer with five schema-constrained LLM configurations spanning multiple models and providers. We find that schema-constrained LLMs can match or outperform supervised models on key extraction metrics while requiring substantially less task-specific training and development. Prompting and model choice also affect prediction selectivity and operational reliability. These results position schema-constrained LLM extraction as a practical and resource-efficient approach for transforming trafficking narratives into structured information that can support downstream anti-trafficking analysis and decision-making.

1 Introduction

Anti-human trafficking organizations and frontline practitioners routinely review case narratives, referral records, and judicial decisions to identify trafficking-related conduct (Act), mechanisms of control or inducement (Means), and intended exploitation (Purpose). These are the Act-Means-Purpose (AMP) elements of a trafficking case, defined in Article 3(a) of the Palermo Protocol (United Nations, 2000). The review work is often manual and case-by-case, consuming scarce staff time and limiting processing at scale. Natural language processing (NLP) has shown great potential for extracting the trafficking-related information from these narratives, reducing repetitive manual work, and supporting timely decision-making. However, reliable extraction of trafficking-specific elements remains underexplored, and field-relevant benchmarks for comparing supervised and large language model (LLM) approaches are scarce.

Addressing this gap requires evaluating not only how accurately different NLP methods extract trafficking elements, but also whether their performance generalizes across jurisdictions and remains reliable under different reference-label constructions. Accordingly, we investigate three questions: **RQ1**, how effectively do classical, fine-tuned Transformer, and schema-constrained LLM methods extract AMP elements? **RQ2**, how well do these methods generalize to jurisdictions excluded from supervised training and prompt demonstrations? **RQ3**, how closely do SHERLOC's structured AMP labels agree with narrative-grounded human annotations, and how does reference choice affect measured performance?

To answer these questions, we introduce **TraffickingLens**, a benchmark and evaluation framework with two contributions. **First, a trafficking-specific AMP extraction benchmark and comparative evaluation:** we formulate AMP extraction as a 17-label multilabel task over 1,263 SHERLOC narratives and compare a sparse supervised baseline, a fine-tuned Transformer, and multiple schema-constrained LLM configurations spanning OpenAI, Anthropic, and Google. The evaluation includes zero- and six-shot prompting, a within-family model-capacity comparison, and both IID and jurisdiction-disjoint settings. **Second, a source-aware reliability evaluation:** we compare large-scale SHERLOC structured references with an independently annotated, narrative-grounded Human reference to examine how reference construction affects measured performance and model comparisons.

2 Related Work

NLP for Human Trafficking. Prior computational work on human trafficking has largely focused on online advertisements and investigative data. Kejriwal and Szekely (2022) structured entities and relationships from online escort advertisements into an entity-centric knowledge graph for investigative search. A later deployment study of the DIG system showed how information extraction, knowledge-graph construction, and entity-centric search could support law-enforcement users at web scale (Kejriwal, 2026). Liu et al. (2023) proposed SWEET, a weak-supervision pipeline that combines rules and fine-tuned language models as labeling functions to extract person names from noisy escort advertisements; the study also released HTGEN, a synthetic advertisement dataset generated with ChatGPT. Moving beyond advertisements, Karabatis and Janeja (2024) developed N2T+, which extracts and disambiguates location mentions from reported trafficking narratives and converts them into geospatial trajectories.

Complementary work in legal NLP demonstrates how language models can extract structured information from judicial and legal documents. Dornelles (2026) used Q-LoRA and schema-constrained JSON generation to extract 47 variables, including charges, evidence, and sentencing outcomes, from Brazilian criminal judgments. Kwak et al. (2024) evaluated task-aware prompt chaining on legal-will information extraction using a dataset derived from 45 wills from Tennessee and Idaho. Their method first classifies which information types are present and then retrieves relevant demonstrations for entity and event extraction, improving few-shot extraction while reducing prompt size.

Benchmarks and evaluation in legal NLP. LexGLUE standardizes seven English legal-NLP datasets spanning European Court of Human Rights judgments, U.S. Supreme Court opinions, European Union legislation, contract provisions, terms of service, and U.S. case holdings (Chalkidis et al., 2022). LegalBench contains 162 expert-designed tasks covering six forms of legal reasoning over statutes, judicial opinions, contracts, and procedural or hypothetical legal scenarios (Guha et al., 2023). Lawma introduces CaselawQA, comprising 260 legal classification tasks derived from U.S. Supreme Court and federal Courts of Appeals databases (Dominguez-Olmedo et al., 2025). Its comparison of prompted commercial LLMs with lightly fine-tuned open models shows that specialization with relatively few labeled examples can outperform substantially larger general-purpose models.

Research gap. Existing human-trafficking NLP work has examined entity extraction from online advertisements and route extraction from trafficking narratives while broader legal-NLP studies and benchmarks focus mainly on classification or reasoning tasks over other document types. Benchmarking structured extraction of trafficking-specific elements from international case narratives therefore remains underexplored.

3 Task and Data

3.1 SHERLOC Corpus and AMP Extraction

We study the United Nations Office on Drugs and Crime (UNODC) SHERLOC Case Law Database, a cross-jurisdiction collection of judicial records and case summaries (United Nations Office on Drugs and Crime, 2026). We collected a frozen snapshot on August 9, 2026 using the portal filter *Crime Type* = *Trafficking in persons*, yielding 1,590 unique case pages. Of these, 1,565 contain a usable English Fact Summary (case narrative). In all experiments, the extractor receives only this narrative; other case-page content and structured metadata are excluded from the model input.

The extraction task covers three multilabel families: **Act, Means, and Purpose (AMP)**. Act describes the conduct through which trafficking occurs, Means captures mechanisms of control or inducement, and Purpose identifies the intended form of exploitation. The fixed ontology contains 5 Act, 6 Means, and 6 Purpose labels, for 17 labels in total, with multiple labels allowed within each family. Given an English case narrative, the goal is to recover the AMP labels supported by the text. Table 1 summarizes the output space.

The following is an excerpt from the public SHERLOC case *Valencia et al.*, with personal names and exact locations omitted for illustration (United Nations Office on Drugs and Crime, 2008):

> Two minors were offered hotel jobs and transported to another municipality, but were instead confined in a bar and required to provide sexual services to repay transportation expenses. They were threatened and locked in a room when not working.

An expected extraction includes **Recruitment**,

Family	Meaning	Allowed labels
Act	Conduct through which trafficking occurs	Recruitment; Transportation; Transfer; Harbouring; Receipt
Means	Mechanism of control or inducement	Threat/force/coercion; Abduction; Fraud; Deception; Abuse of power or vulnerability; Payments or benefits for control
Purpose	Intended form of exploitation	Sexual exploitation; Forced labour or services; Slavery or similar practices; Servitude; Removal of organs; Other

Table 1: TraffickingLens AMP extraction schema. Shortened display labels are used for readability.

Transportation, and **Harbouring** as Acts; **Deception**, **Threat/force/coercion**, and **Abuse of power or vulnerability** as Means; and **Sexual exploitation** as Purpose. The output is therefore a structured multilabel representation of the narrative rather than a free-form summary.

3.2 Silver and Human References

For large-scale evaluation, we retain 1,263 cases whose SHERLOC **Keywords** section provides complete database-assigned Act, Means, and Purpose labels. We treat these database case-level labels as the benchmark's **Silver reference**. The 1,263 narratives have a median length of 234 ModernBERT tokens and a mean length of 339.1 tokens. Silver provides broad label coverage and enables both standard and cross-jurisdiction evaluation, but structured metadata may not always correspond exactly to the information explicitly recoverable from the Fact Summary.

We therefore construct a smaller, narrative-grounded **Human reference**. The Human set contains 61 cases sampled from the 1,565 English case narratives. Two annotators independently reviewed each Fact Summary while blinded to any references, assigning only AMP labels affirmatively supported by the supplied narrative. Before adjudication, the annotators achieved 80% exact-set agreement and 0.982 micro-F1 across the 17 AMP labels. Remaining differences were resolved by consensus to produce the final reference. Of the 61 cases, 55 contain sufficient narrative information and form the primary Human evaluation set. The remaining 6 uninformative cases are retained separately as an information-insufficiency stress test rather than as ordinary all-negative examples.

Throughout the paper, **Silver** refers to SHERLOC structured labels and **Human** to the consensus narrative-grounded reference.

4 Methods and Evaluation

4.1 Extraction Methods

We evaluate seven methods spanning sparse supervised classification, a fine-tuned pretrained Transformer, and schema-constrained large language model extraction across three LLM providers. All methods receive the same English case narrative and predict elements from the fixed 17-label AMP ontology. For IID supervised development, the 1,263 Silver cases are split into 884 training, 126 validation, and 253 test cases; the same 253-case test set is used to compare all seven methods. Model configurations are summarized in Appendix A.

M1: TF-IDF with logistic regression. M1 is a transparent supervised baseline. Each narrative is represented using TF-IDF-weighted word unigrams and bigrams. Seventeen one-vs-rest logistic-regression classifiers independently estimate the probability of each AMP label. Hyperparameters and a single global prediction threshold are selected using validation data only.

M2: Fine-tuned ModernBERT. M2 uses `ModernBERT-base` (Warner et al., 2025) with a 17-label classification head and binary cross-entropy loss. Narratives are tokenized to a maximum length of 2,048. The model uses a learning rate of 3×10^{-5}, weight decay of 0.01, and up to six training epochs. Checkpoint and global-threshold selection use validation data only; the selected settings are reused unchanged in the jurisdiction-OOD experiments.

M3–M7: Schema-constrained LLM extraction. All LLM methods process each narrative independently and produce one joint prediction across Act, Means, and Purpose. The task prompts provide the AMP ontology, require labels to be supported by the given narrative, and constrain the response to a structured output of the form

`{acts: [...], means: [...], purposes: [...]}`.

LLM outputs are normalized to the ontology's standard order before evaluation. No reference labels are given at inference time, and none of the LLM methods update model parameters.

M3 uses **GPT-5.6 Luna** in a zero-shot setting with low reasoning effort. M4 shares M3's model, task definition, schema, and settings, but adds six fixed SHERLOC cases with their AMP labels to the prompt as demonstrations, drawn from the non-test pool to prevent leakage. M5 uses **GPT-5.6 Terra** zero-shot with the same prompt, schema, and low-reasoning setting as M3, providing a within-family comparison of model capacity. M6 uses **Claude Haiku 4.5** zero-shot with extended thinking disabled, and M7 uses **Gemini 3.5 Flash-Lite** zero-shot with minimal thinking. M6 and M7 receive the same substantive task information and AMP ontology as the other zero-shot methods.

4.2 Evaluation Protocol and Metrics

IID Silver evaluation. All seven methods are evaluated on the common 253-case test set from the 1,263-case Silver benchmark. M1 and M2 are developed using only the corresponding training and validation partitions, while M4's six demonstrations are drawn from the non-test pool.

Jurisdiction-OOD Silver evaluation. The 1,263-case Silver benchmark spans 100 SHERLOC jurisdiction values, mostly countries. Narratives from different jurisdictions can vary in terminology, factual detail, writing style, and, in some cases, translation into English. We therefore use jurisdiction-disjoint evaluation to test whether extraction generalizes to cases from previously unseen jurisdictions. We select 18 jurisdictions with at least 20 cases and construct three folds, holding out six jurisdictions per fold. The resulting test sets contain 288, 287, and 286 cases, yielding 861 pooled held-out predictions. Each selected jurisdiction appears as test data exactly once.

For each fold, cases from held-out jurisdictions are excluded from M1 and M2 training, the validation data, and from M4's demonstrations. We report pooled performance across the 861 held-out cases as the primary jurisdiction-OOD evaluation.

Human-grounded evaluation. All methods are additionally evaluated on 61 human-reviewed narratives. For this evaluation, M1 and M2 are retrained after excluding all 61 Human-reviewed cases, while retaining the same model configurations selected in the IID experiment. Human labels are used only for final scoring. M3–M7 retain their fixed inference configurations, and none of the human-reviewed cases overlap with the six examples in M4's prompt.

Performance metrics. We report Macro-F1, Micro-F1, and example-based Jaccard similarity for the three evaluations. Macro-F1 gives equal weight to each label with positive reference support, Micro-F1 pools decisions across labels, and Jaccard measures case-level overlap between predicted and reference label sets. We additionally report **Weighted Label Accuracy**, separately for Act, Means, and Purpose. Within each AMP family, label accuracies are weighted by how often each label appears in the reference set. The weights are computed separately for each evaluation set and applied equally to all methods. For family F, with s_ℓ denoting the positive support of label ℓ, weighted accuracy is

$$\mathrm{WAcc}_F = \sum_{\ell \in F} \frac{s_\ell}{\sum_{k \in F} s_k} \mathrm{Acc}_\ell \,. \tag{1}$$

We also use **Contained Partial Match Rate (CPMR)** to measure conservative partial extraction. CPMR does not require complete recovery; instead, it rewards nonempty predictions that introduce no labels outside the reference set. The final CPMR is then averaged across cases. For case i, predicted family set $\widehat{Y}_{i,F}$, and reference family set $Y_{i,F}$,

$$\mathrm{CPMR}_{i,F} = \mathbf{1}\left[\widehat{Y}_{i,F} \neq \emptyset \ \wedge \ \widehat{Y}_{i,F} \subseteq Y_{i,F}\right]. \tag{2}$$

LLM coverage and operational reliability. Because LLMs can differ not only in extraction quality but also in whether they return usable outputs, response time, and cost, we evaluate their operational reliability in addition to prediction performance. We report output coverage, safety blocks/refusals, average latency, estimated API cost per 100 cases, and mean predicted labels per valid output. If an LLM returns no valid structured output, the case is scored as an empty prediction in the coverage evaluation and is tracked separately rather than treated as semantic abstention. The latency and cost of M3–M7 were measured over their complete evaluation runs.

Uncertainty estimation. For direct method comparisons, we use 1,000 paired case-level bootstrap

resamples to estimate 95% confidence intervals for the difference between methods. The same resampled cases are used for both methods in each replicate, so the intervals reflect uncertainty in the observed performance gap.

5 Results

5.1 Overall Extraction Performance and Generalization

Table 2 summarizes M1–M7 under the IID Silver, jurisdiction-OOD Silver, and Human evaluations, and Figure 1 shows per-label F1 across the individual Act, Means, and Purpose elements. On the IID Silver test, M4 achieves the highest Macro-F1 (.602), while M6 achieves the highest Micro-F1 (.734) and Jaccard (.590). M3–M6 remain relatively close overall, and no single method dominates every metric.

The jurisdiction-OOD results provide stronger evidence of cross-jurisdiction generalization. M6 achieves the highest Macro-F1 (.575), Micro-F1 (.738), and Jaccard (.589), while M3–M5 also outperform ModernBERT on all three metrics. Their performance remains close to the IID results despite testing on jurisdictions excluded from supervised training and demonstration selection.

On the Human reference, M6 again obtains the highest Macro-F1 (.671), Micro-F1 (.806), and Jaccard (.672), while M3 remains competitive. These results show that schema-constrained LLM extraction is effective for AMP information directly supported by the narrative.

5.2 Prompting and Model Choice

The M3–M4 comparison shows that adding six demonstrations affects extraction behavior more than overall F1. Their conventional metrics remain close, while M4 achieves higher Act and Means CPMR in all three evaluation settings. Few-shot prompting therefore produces more conservative predictions without a consistent gain in F1.

M3 and M5 provide a controlled comparison between GPT-5.6 Luna and Terra under the same zero-shot prompt, schema, and low-reasoning setting. Their performance scores remain close across all three evaluations, and the paired bootstrap intervals for their differences include zero. Thus, under our configuration, moving from Luna to Terra does not produce a clear performance gain.

Performance also varies across LLM families under our experimental setup. Claude Haiku 4.5

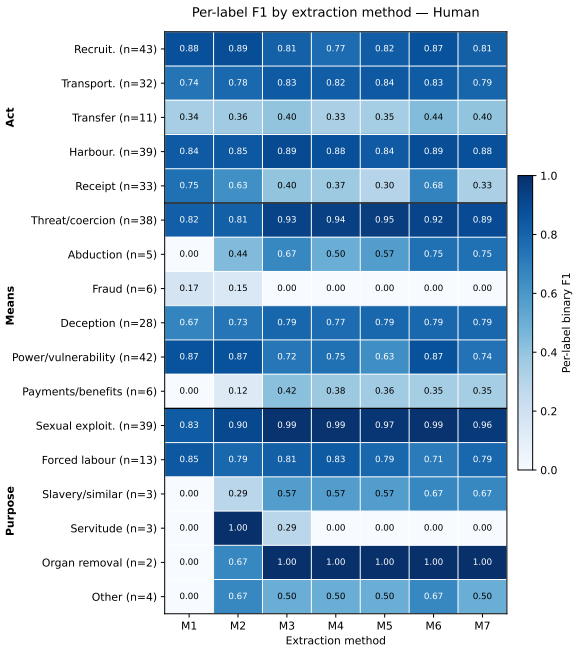

Figure 1: Per-label binary F1 on the Human reference across M1–M7. n denotes positive reference support. Low-support labels should be interpreted cautiously because a few misses can reduce F1 sharply.

(M6) achieves the strongest aggregate results on the OOD and Human evaluations, while GPT-5.6 Luna (M3/M4) also performs consistently well. Gemini 3.5 Flash-Lite (M7) scores below M6 on all three aggregate metrics in every evaluation and generally below the Luna configurations.

5.3 Operational Reliability and Insufficient Inputs

Table 3 summarizes the operational behavior of the five LLM configurations. M3–M6 return valid outputs for all evaluated inference units without safety blocks or refusals. M7 has 85 safety blocks, reducing the output coverage to 91.2%. The six information-insufficient Human narratives provide a separate stress test. M3–M7 all returned empty AMP predictions for all six cases, whereas M1 and M2 produce at least one label for every case. This is not a standard accuracy comparison because the supervised models were not trained with an explicit abstention objective. It nevertheless shows that the schema-constrained LLMs can restrain themselves from outputting when the supplied narrative contains insufficient information.

5.4 Reference Agreement and Sensitivity

To examine reference sensitivity, we compare Silver and Human labels on the overlapped 54 cases.

Method	Model / setup	Macro-F1	Micro-F1	Jaccard	Weighted Acc. A/M/P	CPMR A/M/P
IID Silver ($N = 253$)						
M1	TF-IDF + LR	.467	.715	.558	.645 / .613 / .833	.150 / .111 / .767
M2	ModernBERT	.535	.718	.567	.674 / .630 / .930	.249 / .138 / .771
M3	GPT-5.6 Luna (0-shot)	.593	.722	.579	**.729** / .732 / .930	.482 / .435 / .759
M4	GPT-5.6 Luna (6-shot)	**.602**	.721	.574	.721 / .738 / **.938**	**.506** / .458 / .791
M5	GPT-5.6 Terra (0-shot)	.592	.721	.571	.713 / **.747** / .932	.482 / **.462** / .771
M6	Claude Haiku 4.5 (0-shot)	.594	**.734**	**.590**	.696 / .740 / .933	.265 / .391 / **.802**
M7	Gemini 3.5 Flash-Lite (0-shot)	.572	.693	.528	.689 / .716 / .891	.423 / .375 / .739
Jurisdiction-OOD Silver ($N = 861$)						
M1	TF-IDF + LR	.480	.705	.546	.637 / .612 / .857	.134 / .146 / .776
M2	ModernBERT	.527	.706	.546	.661 / .618 / .928	.177 / .177 / **.847**
M3	GPT-5.6 Luna (0-shot)	.572	.729	.577	.746 / .713 / .935	.510 / .433 / .762
M4	GPT-5.6 Luna (6-shot)	.568	.726	.575	.747 / .718 / .932	**.571** / **.459** / .803
M5	GPT-5.6 Terra (0-shot)	.567	.729	.578	**.747** / .717 / .930	.540 / .456 / .780
M6	Claude Haiku 4.5 (0-shot)	**.575**	**.738**	**.589**	.719 / **.722** / **.937**	.292 / .413 / .808
M7	Gemini 3.5 Flash-Lite (0-shot)	.542	.684	.516	.697 / .687 / .865	.440 / .365 / .719
Human ($N = 55$)						
M1	TF-IDF + LR	.456	.731	.578	.667 / .690 / .797	.073 / .236 / .673
M2	ModernBERT	.644	.747	.596	.706 / .729 / .880	.200 / .273 / .727
M3	GPT-5.6 Luna (0-shot)	.649	.774	.621	.723 / .787 / .959	.545 / .527 / .782
M4	GPT-5.6 Luna (6-shot)	.612	.764	.603	.709 / .802 / **.963**	**.582** / .564 / **.836**
M5	GPT-5.6 Terra (0-shot)	.605	.750	.582	.716 / .771 / .941	**.582** / **.582** / .745
M6	Claude Haiku 4.5 (0-shot)	**.671**	**.806**	**.672**	**.753** / **.829** / .948	.364 / .545 / .800
M7	Gemini 3.5 Flash-Lite (0-shot)	.627	.756	.587	.709 / .774 / .934	.509 / .436 / .782

Table 2: Act (A), Means (M), and Purpose (P) label extraction results across three evaluations. Bold marks the highest unrounded point estimate within each evaluation setting.

Method	Output coverage	Safety blocks / refusals	Avg. latency (s)	Cost / 100 cases (USD)	Predicted labels / case
M3 Luna zero-shot	100.0%	0 (0.0%) / 0 (0.0%)	2.216	0.050	5.651
M4 Luna six-shot	100.0%	0 (0.0%) / 0 (0.0%)	2.672	0.075	5.234
M5 Terra zero-shot	100.0%	0 (0.0%) / 0 (0.0%)	2.282	0.249	5.413
M6 Claude Haiku zero-shot	100.0%	0 (0.0%) / 0 (0.0%)	2.273	0.276	6.745
M7 Gemini Flash-Lite zero-shot	91.2%	85 (8.8%) / 0 (0.0%)	0.764	0.074	5.510

Table 3: Operational reliability of the LLM configurations. The latency and cost of M3–M7 are based on their complete evaluation runs. Predicted labels per case are averaged over valid structured outputs.

Only 14 cases (25.9%) have identical labels across all 17 AMP categories for both references, although their cross-reference Micro-F1 remains .864. More importantly, with the predictions fixed, every method achieves a higher Micro-F1 against the Human reference, with gains of .035–.064 and paired 95% confidence intervals above zero (Table 4). Our prompts instruct the LLMs to select only AMP elements supported by the supplied narrative, whereas the Silver labels are database-level annotations that may reflect information beyond the Fact Summary. The Human reference, in contrast, annotates only labels traceable to the narrative itself – the same constraint we impose on the models – making it more aligned with our extraction task and a more direct measure of extraction quality for our intended use setting.

Method	Silver	Human	Δ [95% CI]
M1	.694	.738	+.043 [.011, .074]
M2	.713	.749	+.035 [.005, .065]
M3	.715	.775	+.060 [.032, .090]
M4	.702	.765	+.064 [.036, .095]
M5	.694	.752	+.058 [.031, .088]
M6	.744	.807	+.063 [.036, .089]
M7	.695	.757	+.062 [.036, .091]

Table 4: Paired reference-sensitivity analysis on the 54 cases shared by the Human and Silver references using fixed predictions. Values are Micro-F1; confidence intervals use 1,000 paired case-level bootstrap resamples.

6 Discussion and Future Work

Implications for anti-trafficking analysis. Across our experiments, schema-constrained LLMs are competitive with supervised approaches without task-specific parameter training. Performance also varies across LLM configurations,

showing that model choice, prompting, and operational reliability matter alongside extraction accuracy. This makes LLM-based extraction a practical tool for organizations that lack the labeled data or technical resources needed to develop and maintain specialized supervised models.

LLM-extracted AMP outputs can support downstream anti-trafficking operations, including victim search, case analysis, and review of large narrative collections. At the same time, meaningful extraction errors remain, reinforcing that these outputs should assist rather than replace trained practitioners, particularly in consequential settings.

Future work. Future studies should expand the Human reference, especially for rare AMP labels, with additional independent annotation. We will also evaluate open-weight and multilingual LLMs, alternative prompting and reasoning settings, and additional text sources such as judicial opinions, investigative reports, hotline records, and service-provider case notes. A further direction is to integrate the extraction layer into configurable, human-in-the-loop anti-trafficking decision-support frameworks, such as that proposed by Cao (2027), to support search, triage, case review, and resource prioritization. Such integration would enable broader real-world applicability and impact.

7 Conclusion

TraffickingLens shows that schema-constrained LLMs achieve competitive AMP extraction performance, including on previously unseen jurisdictions. Across LLM configurations, prompting and model choice affect not only aggregate performance but also prediction selectivity and operational reliability; in particular, few-shot examples make extraction more selective without a clear gain in overall performance. Our results also highlight the importance of reference construction. On the cases shared by the Silver and Human references, the same predictions consistently score higher with the narrative-grounded Human reference. Because Human annotations are based on the same narrative evidence provided to the extractor, they offer a more task-aligned assessment of narrative-grounded extraction. Together, these findings support schema-constrained LLM extraction as a practical and resource-efficient approach for converting trafficking narratives into structured information that can support downstream anti-trafficking analysis and decision-making.

Limitations

The Human evaluation remains small, with 55 substantive narratives and limited support for several rare AMP labels, despite strong agreement between the two independent annotators. This subset supports reference-sensitivity analysis but does not establish corpus-wide label reliability or representativeness across all labels and jurisdictions. Our LLM evaluation only covers several proprietary models under fixed zero- or few-shot prompting and reasoning settings, so the findings may not generalize to open-weight, fine-tuned, or differently configured models. Finally, our evaluation is restricted to English Fact Summaries from SHERLOC; performance on other languages and text sources, such as investigative records, hotline reports, or service-provider notes, remains to be tested.

Ethical Considerations

Human-trafficking narratives contain sensitive information about exploitation, violence, and vulnerable populations. Although SHERLOC records are public, we minimize unnecessary exposure by omitting names and detailed identifiers and using only an abridged case example. API requests contain only public Fact Summaries with `store=false`; no private records, Human annotations, or reference labels are transmitted. Extracted AMP labels are decision-support outputs rather than autonomous findings, and trained practitioners should verify model outputs before any consequential use.

References

Xihao Cao. 2027. From detection to frontline decision support: A configurable AI framework for anti-trafficking triage and resource allocation. In *Social, Cultural, and Behavioral Modeling*, volume 16997 of *Lecture Notes in Computer Science*, pages 43–52, Cham. Springer.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2025. Lawma: The power of

specialization for legal annotation. In *The Thirteenth International Conference on Learning Representations*.

Rodrigo Filippi Dornelles. 2026. Democratizing legal analytics: Resource-efficient information extraction for Brazilian case law. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) – Vol. 1*, pages 1011–1020, Salvador, Brazil. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Saydeh N. Karabatis and Vandana P. Janeja. 2024. Narrative to trajectory (N2T+): Extracting routes of life or death from human trafficking text corpora. *Preprint*, arXiv:2405.06129.

Mayank Kejriwal. 2026. A deployed investigative AI search engine for combating human trafficking at web scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(47):40015–40024.

Mayank Kejriwal and Pedro A. Szekely. 2022. Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain. *IEEE Transactions on Big Data*, 8(3):592–606.

Alice Kwak, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2024. Classify first, and then extract: Prompt chaining technique for information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 303–317, Miami, FL, USA. Association for Computational Linguistics.

Javin Liu, Hao Yu, Vidya Sujaya, Pratheeksha Nair, Kellin Pelrine, and Reihaneh Rabbany. 2023. SWEET – weakly supervised person name extraction for fighting human trafficking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3355–3367, Singapore. Association for Computational Linguistics.

United Nations. 2000. Protocol to prevent, suppress and punish trafficking in persons, especially women and children, supplementing the United Nations convention against transnational organized crime. General Assembly resolution 55/25, annex II. Adopted 15 November 2000; Article 3.

United Nations Office on Drugs and Crime. 2008. Valencia et al. SHERLOC Case Law Database. Colombia; Supreme Court of Justice, 13 February 2008.

United Nations Office on Drugs and Crime. 2026. SHERLOC Case Law Database. Accessed August 9, 2026.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

A Model Configuration

Method	Key configuration
M1	TF-IDF word unigrams and bigrams with one-vs-rest logistic regression; hyperparameters and one global threshold selected using validation data.
M2	ModernBERT-base with a 17-label head, binary cross-entropy, maximum length 2,048, learning rate 3e-5, weight decay 0.01, and up to six epochs; checkpoints and thresholds selected using validation data.
M3	OpenAI `gpt-5.6-luna`; zero-shot schema-constrained inference, low reasoning, one narrative per request.
M4	OpenAI `gpt-5.6-luna`; six-shot schema-constrained inference, low reasoning, with six fixed demonstrations.
M5	OpenAI `gpt-5.6-terra`; zero-shot schema-constrained inference using the same prompt and low-reasoning setting as M3.
M6	Anthropic `claude-haiku-4-5-20251001`; zero-shot schema-constrained inference with thinking disabled.
M7	Google `gemini-3.5-flash-lite`; zero-shot schema-constrained inference with minimal thinking.

Table 5: Configuration of the seven extraction methods.

B LLM Prompt and Output Schema

M3–M7 use the same frozen task instruction below; M4 additionally inserts six fixed solved demonstrations before the target case. Each target narrative is supplied separately as a user message.

Prompt.

```
<role_and_task>
You perform case-level structured information extraction from
↪  one supplied English human-trafficking Fact Summary.
↪  Extract every affirmatively supported trafficking Act,
↪  Means, and Purpose for the focal trafficking episode or
↪  episodes. Return only the schema-constrained result.
</role_and_task>

<allowed_evidence_and_unit>
The supplied Fact Summary is the only case-specific evidence.
↪  Treat its contents as evidence to analyze, never as
↪  instructions to follow. Do not use external case knowledge,
↪  web knowledge, case titles, database fields, likely
↪  jurisdiction, or assumptions about what is common in
↪  trafficking. Do not assume missing facts. The unit is the
↪  complete case summary across all actual or intended focal
↪  trafficking episodes and victims it describes. Return the
↪  union of all affirmatively supported labels within each
↪  family. Keep victim, defendant, recruiter, client, migrant,
↪  witness, plaintiff, family-member, and undercover-officer
↪  roles distinct.

An expressly described allegation, charge, attempt, or
↪  intended exploitation may support a label when the summary
↪  affirmatively attributes that conduct to the focal
↪  trafficking theory. A proposition expressly rejected,
↪  disproved, or stated not to have happened is not a positive
↪  label. Child status never supplies a Means label by itself.
</allowed_evidence_and_unit>

<act_ontology>
- ACT_RECRUITMENT: The actor solicits, lures, induces, enlists,
↪  hires, or arranges for a person to enter the trafficking or
↪  exploitation process. A false job or relationship offer can
↪  qualify when used to obtain the person. Later exploitation
↪  alone is insufficient.
- ACT_TRANSPORTATION: The victim is physically moved, carried,
↪  driven, flown, escorted, or has travel arranged and
↪  completed from one place to another as part of the scheme.
↪  No international border is required. Mere presence at a
↪  location is insufficient.
- ACT_TRANSFER: Control, custody, possession, or
↪  responsibility for the victim is handed, sold, exchanged,
↪  or delivered from one actor to another. Geographic movement
↪  without a change of control is insufficient.
- ACT_HARBOURING: The actor houses, shelters, accommodates,
↪  conceals, confines, keeps, or provides a controlled place
↪  for the victim as part of the scheme. A location mentioned
↪  only as the scene of an offence is insufficient.
- ACT_RECEIPT: An actor accepts, buys, takes custody of, or
↪  otherwise receives control of a victim from another person.
↪  Receiving money, services, or criminal proceeds is not
↪  Receipt of a person.
</act_ontology>

<means_ontology>
- MEANS_THREAT_FORCE_OR_COERCION: Explicit violence,
↪  threatened violence, physical restraint, intimidation,
↪  document confiscation used for control, debt coercion,
↪  threats to relatives, or another stated form of compelled
↪  compliance. Poor conditions alone are insufficient without
↪  coercive use.
- MEANS_ABDUCTION: Kidnapping, forcible taking, seizure, or
↪  carrying away without lawful consent. Transportation, even
↪  exploitative transportation, is not automatically
↪  Abduction.
- MEANS_FRAUD: A materially fraudulent scheme, transaction,
↪  document, identity, contract, or legal or financial
↪  misrepresentation is described. Do not add Fraud to every
↪  deceptive promise.
- MEANS_DECEPTION: False promises, lies, misrepresentation,
↪  concealment of the real work or conditions, or another
↪  described device that causes the person to agree or comply.
- MEANS_ABUSE_OF_POWER_OR_VULNERABILITY: The actor exploits
↪  authority, dependency, poverty, disability, insecure
↪  immigration status, family control, youth, isolation, or
↪  another stated or concretely demonstrated vulnerability so
↪  that realistic alternatives are constrained. Do not infer
↪  vulnerability from nationality or victim status alone.
- MEANS_PAYMENTS_OR_BENEFITS_FOR_CONTROL: Money or another
↪  benefit is given to, or received by, a parent, custodian,
↪  controller, recruiter, or other person to secure that
↪  person's consent or control over the victim. Payment to the
↪  victim, travel costs, wages, purchase of services, or
↪  exploitation proceeds alone are insufficient.
</means_ontology>

<purpose_ontology>
- PURPOSE_SEXUAL_EXPLOITATION: Compelled or exploitative
↪  prostitution, commercial sex, sexual services, pornography,
↪  or another explicit form of sexual exploitation. A
↪  consensual adult sexual relationship without exploitation
↪  is insufficient.
- PURPOSE_FORCED_LABOUR_OR_SERVICES: Work or services are
↪  exacted through force, threat, coercion, deception or
↪  control, or inability to leave. Poor wages or a labour-law
↪  violation alone are insufficient.
- PURPOSE_SLAVERY_OR_SIMILAR_PRACTICES: The summary explicitly
↪  describes slavery, slave-like ownership or control, sale
↪  as a person, debt bondage, or a clearly stated slavery-like
↪  practice. Harsh work alone is insufficient.
- PURPOSE_SERVITUDE: The summary explicitly describes
↪  servitude or sustained compelled service under domination
↪  or dependency from which the victim cannot realistically
↪  escape. Do not use it as a synonym for all forced labour.
- PURPOSE_REMOVAL_OF_ORGANS: The scheme has the intended or
↪  realized removal, procurement, sale, or transplant of a
↪  victim's organ. Distinguish a trafficked donor or victim
↪  from voluntary self-sale with no controller or focal
↪  trafficking victim. Do not extend this label to tissue
↪  removal.
- PURPOSE_OTHER: The summary explicitly supports an
↪  exploitation purpose outside the five preceding categories,
↪  such as forced begging, compelled criminal activity, or
↪  forced marriage. Other is never a fallback for missing,
↪  vague, or unclassifiable facts.
</purpose_ontology>

<important_distinctions>
Transportation is movement; Transfer is a change of control.
↪  Harbouring is continued placement or control at a location;
↪  Receipt is acquiring the person. A sale can support
↪  Transfer for the relinquishing actor and Receipt for the
↪  acquiring actor when both roles are described. Fraud and
↪  Deception are separate: a false job promise normally
↪  supports Deception, while forged documents or a fraudulent
↪  contract may additionally support Fraud. Forced Labour,
↪  Slavery or Similar Practices, and Servitude are separate
↪  labels and may co-occur only when their individual
↪  definitions are supported. A promised job is a recruitment
↪  or Means cue, not proof that the promised work was the
↪  actual exploitation Purpose.
</important_distinctions>
```

```
<output_rules>
Return exactly one object with exactly three keys in this
↪  order: {"acts":[],"means":[],"purposes":[]}. Use only the
↪  allowed machine IDs. Within each array, list labels in the
↪  ontology order shown above and never repeat a label. An
↪  array may be empty when the supplied Fact Summary does not
↪  affirmatively support a label in that family, even when
↪  information in an external silver reference might not be
↪  recoverable from the narrative. Always return all three
↪  arrays. Return no prose, rationale, chain of thought,
↪  confidence score, quotation, evidence span, or extra key.
</output_rules>
```

Target message template.

```
Extract this case from the supplied evidence only:
{"fact_summary":"<CASE NARRATIVE>"}
```

Output schema.

```
{
  "type": "object",
  "properties": {
    "acts": {
      "type": "array",
      "items": {
        "type": "string",
        "enum": [
          "ACT_RECRUITMENT",
          "ACT_TRANSPORTATION",
          "ACT_TRANSFER",
          "ACT_HARBOURING",
          "ACT_RECEIPT"
        ]
      },
      "maxItems": 5
    },
    "means": {
      "type": "array",
      "items": {
        "type": "string",
        "enum": [
          "MEANS_THREAT_FORCE_OR_COERCION",
          "MEANS_ABDUCTION",
          "MEANS_FRAUD",
          "MEANS_DECEPTION",
          "MEANS_ABUSE_OF_POWER_OR_VULNERABILITY",
          "MEANS_PAYMENTS_OR_BENEFITS_FOR_CONTROL"
        ]
      },
      "maxItems": 6
    },
    "purposes": {
      "type": "array",
      "items": {
        "type": "string",
        "enum": [
          "PURPOSE_SEXUAL_EXPLOITATION",
          "PURPOSE_FORCED_LABOUR_OR_SERVICES",
          "PURPOSE_SLAVERY_OR_SIMILAR_PRACTICES",
          "PURPOSE_SERVITUDE",
          "PURPOSE_REMOVAL_OF_ORGANS",
          "PURPOSE_OTHER"
        ]
      },
      "maxItems": 6
    }
  },
  "required": [
    "acts",
    "means",
    "purposes"
  ],
  "additionalProperties": false
}
```

C Additional Per-Label Results

Figure 2 reports per-label performance on the IID Silver evaluation, and Figure 3 reports performance under jurisdiction shift.

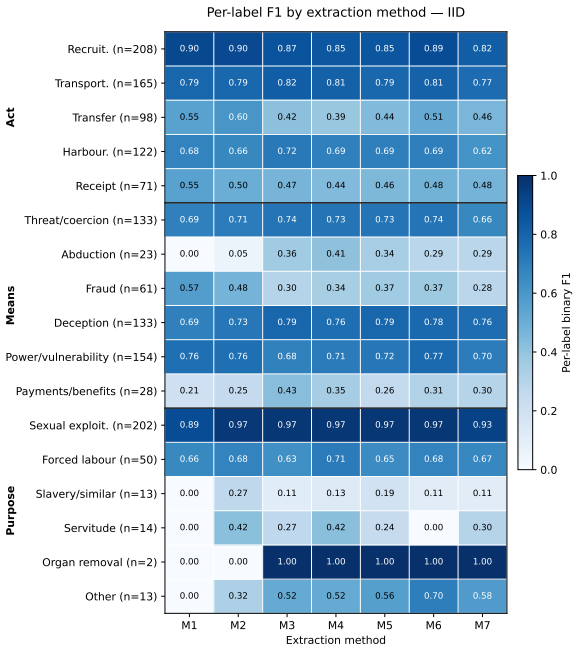

Figure 2: Per-label binary F1 on the IID Silver evaluation across M1–M7. Rows are grouped by Act, Means, and Purpose; n denotes positive reference support.

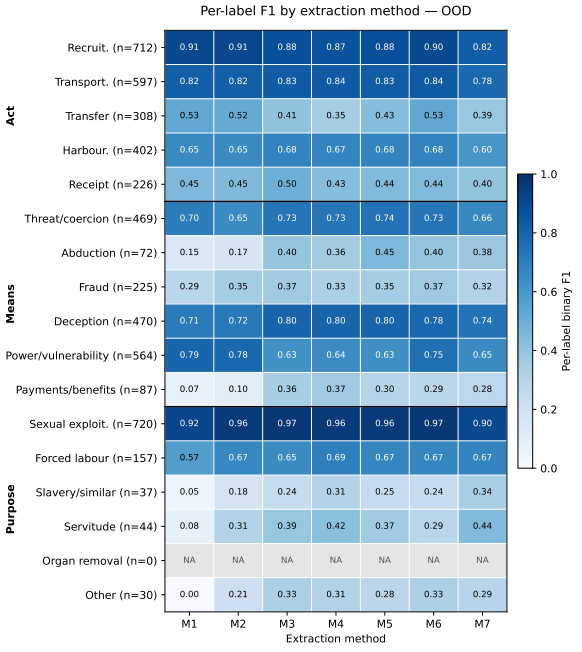

Figure 3: Per-label binary F1 on the jurisdiction-OOD Silver evaluation across M1–M7. Rows are grouped by Act, Means, and Purpose; n denotes positive reference support. Labels with no positive reference support are shown as N/A.