Open-Domain Safety Policy Construction

Di Wu[1], **Siyue Liu**[1], **Zixiang Ji**[1], **Ya-Liang Chang**[2], **Zhe-Yu Liu**[2], **Andrew Pleffer**[2], **Kai-Wei Chang**[1]
[1]University of California, Los Angeles [2]Taboola
{diwu, kwchang}@cs.ucla.edu

Abstract

Moderation layers are increasingly a core component of many products built on user- or model-generated content. However, drafting and maintaining domain-specific safety policies remains costly. We present **Deep Policy Research** (DPR), a minimal agentic system that drafts a full content moderation policy based on only human-written seed domain information. DPR uses a single web search tool and lightweight scaffolding to iteratively propose search queries, distill diverse web sources into policy rules, and organize rules into an indexed document. We evaluate DPR on (1) the OpenAI undesired content benchmark across five domains with two compact reader LLMs and (2) an in-house multimodal advertisement moderation benchmark. DPR consistently outperforms definition-only and in-context learning baselines, and in our end-to-end setting it is competitive with expert-written policy sections in several domains. Moreover, under the same seed specification and evaluation protocol, DPR outperforms a general-purpose deep research system, suggesting that a task-specific, structured research loop can be more effective than generic web research for policy drafting. We release our experiment code at https://github.com/xiaowu0162/deep-policy-research.

1 Introduction

Content moderation modules are core layers in modern products for managing unsafe or low-quality inputs. These systems are guided by domain-specific policies that define allowed and disallowed content and support consistent enforcement across labeling, training, and deployment (Markov et al., 2023; Vidgen and Derczynski, 2020; Yin and Zubiaga, 2021; Zeng et al., 2020). However, drafting and maintaining high-quality policies remains costly. It requires domain expertise, repeated iteration, and frequent updates as products evolve and new edge cases appear. While studies have explored automatic pipelines to improve the effectiveness or reduce the cost of applying the policies Markov et al. (2023); Guan et al. (2024), a fully human-written policy is still a prerequisite. In this paper, we challenge this assumption by asking:

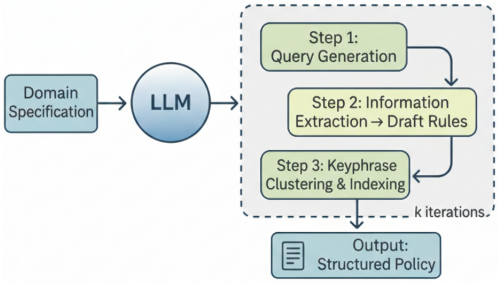

Figure 1: An illustration of Deep Policy Research. Based on a domain specification, an LLM iteratively interacts with a search engine, extracts policy rules, and indexes the rules through keyphrase-based clustering.

Can we leverage LLMs to assist in drafting the policies themselves?

To begin with, we frame the task **open-domain policy construction**. The input is a concise domain specification and access to a search engine. The output is a structured policy document. Success is measured by downstream utility, such as the accuracy of a fixed moderation model when the policy is provided in-context.

We then propose **Deep Policy Research** (DPR). DPR is a minimal agent that uses only web search as an external tool and a lightweight scaffolding scheme for rule writing. Starting from a one-sentence domain definition, DPR iteratively identifies missing coverage, issues targeted queries, distills retrieved sources into rule predicates, and consolidates them into an indexed policy document (Figure 1). The indexing stage organizes rules into coherent sections, improving readability and helping reader models consume long policies.

Findings of the Association for Computational Linguistics: EACL 2026, pages 4981–4999

We evaluate DPR in two content moderation settings. On the OpenAI undesired content benchmark (Markov et al., 2023) across five domains, DPR improves moderation F_1 over both definition-only prompting and few-shot in-context examples for two compact reader LLMs. Averaged across domains, DPR increases F_1 from 0.752 to 0.792 on Llama 3.1 8B and from 0.810 to 0.831 on Qwen2.5 7B, with the largest gains on more subjective categories such as Violence, Harassment, and Self-Harm (§4.2). Under the same seed specification and evaluation protocol, DPR also outperforms a general-purpose deep research system, improving average F_1 from 0.776 to 0.792 on Llama 3.1 8B and from 0.800 to 0.831 on Qwen2.5 7B (§4.2). We further evaluate DPR on an in-house multimodal advertisement moderation benchmark. Replacing an expert-written domain section with a DPR-generated section recovers much of the human policy benefit in several domains and substantially improves over removing the section or using only the one-sentence specification (§5.2).

In summary, we introduce open-domain policy construction and evaluate it by downstream utility. We propose Deep Policy Research, a minimal agent that uses web search and lightweight scaffolding to synthesize and index domain-specific policies. Across text-only and multimodal moderation settings, DPR-generated policies improve downstream moderation and are competitive with expert-written policy sections in several domains, while outperforming a general-purpose deep research baseline under the same protocol. DPR also provides a reproducible environment for future work on policy drafting agents, including reader-model-specific policy presentation and generating illustrative examples alongside rule predicates to clarify decision boundaries.

2 Related Work

Policy Use in LLM Systems Recent alignment strategies explicitly incorporate human-written safety policies into the training or reasoning process of large language models. Deliberative Alignment fine-tunes LLMs to reason based on an entire written policy before responding, yielding safer outputs that strictly adhere to guidelines (e.g. reduced jailbreaks and fewer unjustified refusals) (Guan et al., 2024). Anthropic's Constitutional AI similarly forgoes direct human feedback in favor of a fixed set of normative principles that the model internalizes as a "constitution," using them to self-criticize and refine its answers (Bai et al., 2022). Other works integrate policies as part of the reward or decision mechanism: for instance, OpenAI's GPT-4 alignment process included a rule-based reward model that penalized policy violations during RLHF (Mu et al., 2024), and recent safety reasoning frameworks train models to follow chain-of-thought traces grounded in explicit policy rules (Mou et al., 2025).

Policy Writing The direction of automatically creating or refining safety policies has been less explored. OpenAI has demonstrated that GPT-4 can assist policy designers by identifying ambiguities and edge cases in draft guidelines: the model labels content according to a given policy, explains any discrepancies with human judgments, and suggests clarifications, thereby accelerating the policy refinement loop (OpenAI, 2023). We build upon this intuition and take a step forward, automatically generating policy end-to-end from human-curated domain specifications.

3 Approach

3.1 Problem Formulation

We study **open-domain policy construction**. The input is a domain specification s that describes scope and intent for a single moderation domain, and a search engine $\mathcal{G}$. The output is a policy document P consisting of a set of textual rules organized into sections. P can be free-formed or following a hierarchical structure. We evaluate P by **downstream utility** in a fixed content moderation setup, where a reader LLM receives P in-context and performs safe vs. unsafe binary classifications. A policy construction system succeeds if its output policy improves downstream moderation performance under the same reader model and evaluation protocol.

3.2 Deep Policy Research

We present **Deep Policy Research** (DPR), a minimal research agent that constructs a policy by iteratively searching the web and distilling sources into structured rules. DPR uses an LLM $\mathcal{M}$ as the research model and web search $\mathcal{G}$ as the only external tool. DPR runs for k iterations and maintains two artifacts at iteration i: a policy draft P_i and an index I_i that organizes rules into sections. DPR initializes $P_0 \equiv s$ and $I_0 \equiv s$.

Concretely, at each iteration $i \in \{1, \ldots, k\}$, DPR performs three steps.

Step 1: Query generation. DPR first analyzes the current policy organization I_{i-1} and proposes a set of research queries Q_i to expand coverage or refine ambiguous parts of the policy. Queries are written to target definitional boundaries, common edge cases, high-risk subtypes, and enforcement cues. For each query $q \in Q_i$, DPR retrieves the top m search results using $\mathcal{G}$ and collects page titles, snippets, and URLs as evidence for rule extraction.

Step 2: Rule extraction and consolidation. Given the retrieved evidence, DPR prompts $\mathcal{M}$ to extract candidate rules in a consistent schema. Each rule is written as a short predicate-style statement with a clear decision boundary and optional qualifiers. Concretely, we ask $\mathcal{M}$ to produce rules that specify a condition and a moderation decision, and to include brief scope qualifiers when needed. DPR then runs a self-critique pass to improve precision and reduce noise. In this pass, $\mathcal{M}$ removes irrelevant or overly generic rules, merges redundant rules that express the same decision boundary, and resolves conflicts by preferring rules that are supported by multiple sources or by higher-quality sources. The output is a consolidated rule set R_i.

Step 3: Indexing. DPR merges the new rules into the policy draft, $P_i \leftarrow P_{i-1} \cup R_i$. It then organizes the full rule set into sections to form an indexed policy document I_i. We use keyphrase-based clustering to build this index. DPR asks $\mathcal{M}$ to extract keyphrases for each rule, clusters keyphrases into n groups using k-means, and asks $\mathcal{M}$ to name each cluster and write a short section summary that captures the shared theme. Finally, DPR merges clusters with overlapping semantics to produce a compact, readable index. The resulting I_i is used both as a human-readable policy document and as a coverage signal for the next query generation step.

After k iterations, DPR outputs the final indexed policy $P \equiv I_k$. Figure 1 illustrates the overall loop. Our goal is not to engineer complex agent architectures or elaborate scaffolding strategies. Instead, we aim to isolate a minimal, reproducible research loop that uses a single external tool and lightweight human scaffolding, and to show that this simple design can already produce useful, structured policies with measurable downstream utility. For reproduction, we list the domain specifications in §A.3 and the prompts in §A.4.

4 Results: OpenAI Content Moderation

4.1 Experimental Setup

We evaluate DPR on the OpenAI undesired content benchmark introduced in Markov et al. (2023). Following prior work, we consider five major moderation domains and report binary classification F_1 computed per domain and averaged across domains. In all settings, we keep the downstream reader LLM fixed and vary only the policy provided in-context, so differences reflect the utility of the constructed policy rather than changes in the classifier or model weights. Full implementation details are provided in §A.1.

Baselines We compare DPR with three baselines:

- **Seed Information**: judging with only the seed information s_i for each domain.
- **In-Context Learning**: we randomly sample three unsafe examples and three safe examples as the in-context demonstrations.
- **OAI DR**: we manually run the OpenAI Deep Research Agent through the WebUI. The seed information is provided and we use GPT-5.1 as the research model.

4.2 Content Moderation Accuracy

Table 1 reports results across five domains and two reader LLMs. DPR consistently improves over both Seed Information and In-Context Learning for every domain and reader model. Averaged across domains, DPR increases F_1 from 0.752 to 0.792 on Llama 3.1 8B and from 0.810 to 0.831 on Qwen2.5 7B. Gains are most pronounced in more subjective categories such as Violence, Harassment, and Self-Harm, where definition-only prompts leave substantial ambiguity and few-shot demonstrations are brittle. Notably, DPR does not introduce regressions on well-specified categories. For Sexual, differences are within 0.01 F_1 for both readers, indicating that web-grounded refinement can add coverage without injecting noise.

Under the same seed specification and evaluation protocol, DPR also outperforms the general-purpose deep research baseline. On Llama 3.1 8B, the average F_1 improves from 0.776 with OAI DR to 0.792 with DPR. On Qwen2.5 7B, the average improves from 0.800 to 0.831. This suggests that task-specific structure matters for policy drafting:

	Sexual	Hate	Violence	Harassment	Self-Harm	Average
`Llama 3.1 8B Instruct`						
Seed Information	0.916	0.757	0.658	0.640	0.788	0.752
In-Context Learning	0.923	0.728	0.582	0.610	0.691	0.707
OAI DR	0.828	0.800	0.738	0.683	0.829	0.776
DPR	0.910	0.813	0.717	0.683	0.835	**0.792**
`Qwen2.5 7B Instruct`						
Seed Information	0.939	0.817	0.782	0.670	0.842	0.810
In-Context Learning	0.927	0.803	0.645	0.551	0.779	0.741
OAI DR	0.919	0.765	0.763	0.781	0.773	0.800
DPR	0.949	0.811	0.784	0.752	0.860	**0.831**

Table 1: Evaluation results on OpenAI Content Moderation. DPR improves content moderation F1, outperforming both only using seed information or performing in-context learning with human-written examples.

a simple loop that enforces rule extraction, consolidation, and indexed organization can yield more usable policies than generic web research when the end goal is downstream moderation.

Overall, these results support two conclusions. First, open-domain policy construction can translate into measurable moderation gains even when the downstream model is held constant. Second, a minimal agent with a single external tool can be effective when paired with lightweight human scaffolding that constrains outputs into actionable rule predicates and an indexed document. We provide additional analyses on the research model and indexing design in §B. Further qualitative examples are provided in §B.3.

5 Results: In-house Multimodal Advertisement Moderation

5.1 Experimental Setup

We further evaluate DPR in a real-world multimodal moderation setting using an in-house advertisement benchmark. Each example consists of ad text and a thumbnail image, and the task is to predict whether the creative complies with an in-house safety policy. We report binary classification F_1 and evaluate with a fixed vision-language reader model that receives the policy in-context. We consider two inference settings, single-sample decoding (S.S.) and majority voting (M.V.) over multiple samples. Full dataset and evaluation details are provided in §A.2.

This benchmark includes a comprehensive human-written policy document, where each section corresponds to a policy domain. Our evaluation therefore focuses on **substituting** a single domain section while keeping the remainder of the policy fixed. We compare five configurations:

1. **No Policy** removes the target domain section entirely while keeping the other sections.
2. **Human Policy** uses the full original expert-written policy document.
3. **Seed Information** replaces the section with the one-sentence domain specification.
4. **DPR** replaces the section with a DPR-generated policy built from the same one-sentence specification and web search.
5. **DPR + Summary** further compresses the DPR index into a shorter rule set to reduce prompt length.

5.2 Content Moderation Accuracy

Table 2 reports results for four domains under both inference settings. Removing the domain section or replacing it with only the one-sentence specification degrades performance in three of four domains, showing that the domain-specific section carries substantial utility beyond the rest of the policy. In contrast, substituting the section with DPR recovers much of this benefit. Under single-sample inference, DPR improves the average F_1 from 0.68 with No Policy and 0.69 with Seed Information to 0.75. Gains are largest for visually nuanced categories such as Exploitative and Offensive, where the web-sourced rules often capture concrete cues and common edge cases that are missing from a short specification.

The fully human-written policy remains strongest overall, but DPR narrows the gap substantially in several domains. Under majority voting, DPR reaches near parity on Misrepresentative, achieving 98% of the human policy performance, and it comes within a small

	Misrepresentative		Finance Claims		Exploitative		Offensive	
	S.S.	M.V.	S.S.	M.V.	S.S.	M.V.	S.S.	M.V.
No Policy	0.714	0.727	0.500	0.509	0.793	0.782	0.714	0.786
Seed Information	0.701	0.701	0.553	0.544	0.839	**0.885**	0.679	0.786
Human Policy	**0.740**	**0.779**	**0.833**	**0.877**	**0.920**	0.908	**0.893**	**0.964**
DPR	0.727	0.740	0.597	0.597	0.908	**0.920**	0.786	0.821
DPR + Summary	0.701	**0.779**	0.588	0.614	0.874	0.908	0.821	0.821

Table 2: Evaluation results on in-house multimodal moderation data. DPR improves content moderation F1, outperforming both no policy and only using seed information. DPR also performs on par with human-written policy in the Mispresentative and Exploitative domain. We provide further human evaluations in Appendix §C.

margin on Exploitative. At the same time, Finance Claims remains a clear outlier. This category relies heavily on organization-specific compliance language and fine-grained disclaimer patterns, which are difficult to recover from open web sources alone. These results highlight both the promise and limits of open-domain policy construction. DPR can quickly bootstrap useful draft sections from minimal input, especially in domains where conventions and edge cases are well represented in public guidance, while expert-written rules remain important for categories driven by proprietary standards.

Finally, DPR + Summary reduces the policy length but can trade off accuracy. The compressed policy retains most of the gains on Exploitative and Offensive but loses 1–2 F_1 on Finance Claims and Misrepresentative, suggesting that aggressive compression can remove rare qualifiers that the reader model uses for borderline decisions. We present qualitative examples of the generated rules and their relationship to expert policy sections in §B.3. We also provide a human evaluation of the generated rules in §C.

6 Conclusion

Deep Policy Research (DPR) shows that open-domain research agents can autonomously synthesize web-sourced information into effective safety policies. Across online text moderation and multimodal ad review, DPR outperforms definition-only prompts and example-based learning and often recovers nuanced edge cases and policy conventions, demonstrating the feasibility of drafting policies from high-level human guidance.

Limitations

Despite these promising results, several directions remain open for future work. First, beyond textual rules, future systems could synthesize representative safe and unsafe examples to provide more actionable guidance. Such example-driven policies may improve downstream interpretability and generalization. Second, while DPR currently relies solely on web content, incorporating human feedback during rule generation or filtering could improve factuality, alignment with organizational values, and trustworthiness. Finally, although we focus on content moderation, DPR's methodology may generalize to broader safety-critical applications such as alignment training data curation, AI deployment policies, or auditing frameworks. We hope our work motivates further exploration into autonomous, scalable policy generation methods that complement human oversight and accelerate the development of safer AI systems.

Ethics Statements

Potential Risks Our goal is to study an agentic framework (DPR) that drafts safety policies by researching the open web. Even though the downstream task is safety-oriented, the approach could introduce new risks. First, web-sourced material may encode historical or societal biases. DPR could surface, compress, or over-generalize such biases into policy text, which may propagate into model decisions. Second, policy synthesis can fail subtly (e.g., conflating jurisdictions, misreading sources, or hallucinating "best practices"), which could lead to over-blocking (chilling legitimate speech) or under-blocking (missing harmful content). To mitigate these risks, we recommend that users approach DPR outputs as advisory drafts rather than definitive policy, retain human oversight from relevant reviewers, and plan to revisit decisions as needs, data, and community expectations evolve.

Artifact Creation and Usage Meta Llama 3.1/3.3 Instruct are under the Llama Community License (with Acceptable Use Policy). We use them for research benchmarking only and do not redistribute weights. Qwen2.5-7B-Instruct is under the Apache-2.0 license. We use it for benchmarking and ablations with required notices preserved. The OpenAI API model (e.g., GPT-4o) is governed by OpenAI's Terms of Use. We use it as an evaluation model/LLM-as-judge through the API without redistributing model artifacts. The OpenAI "Undesired Content" evaluation dataset is under the MIT License. We use it strictly for research benchmarking with attribution, consistent with MIT terms. Our in-house multimodal advertisement benchmark is proprietary. We use it only to compute and report aggregate metrics and do not release underlying creatives. Our DPR code is released under MIT. We provide it for replication and extension. Our DPR-generated policy texts and indices are released under MIT. We will provide them for research and educational reuse with attribution and without commercial use. We may cite short excerpts from third-party web pages for context with attribution and do not redistribute full texts.

AI Assistant Use AI assistants, specifically ChatGPT, are used only for improving the paper writing and the appearance of the figures.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2024. Deliberative alignment: Reasoning enables safer language models. *CoRR*, abs/2412.16339.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 15009–15018. AAAI Press.

Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. 2025. SaRO: Enhancing LLM safety through reasoning-based alignment. *arXiv preprint arXiv:2504.09420*.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*.

OpenAI. 2023. Using GPT-4 for content moderation. `https://openai.com/index/using-gpt-4-for-content-moderation/`. Accessed 2025-07-04.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *ArXiv preprint*, abs/2412.15115.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2020. Bad news: Clickbait and deceptive ads on news and misinformation websites. In *Workshop on Technology and Consumer Protection (ConPro)*.

Supplementary Material: Appendices

A Implementation Details

A.1 OpenAI Content Moderation

Evaluation We select the five major domains (sexual, hate, violence, harassment, and self-harm) and use their official definition as the input for DPR[1], as presented in Figure 2. For each domain, we reserve five examples for validation and use the rest for testing. We perform evaluation separately for each domain using the in-domain undesired test examples as well as the safe test examples. To define the utility metric for the system-generated policy, we focus the evaluation scope on long-context evaluation with an off-the-shelf LLM. In other words, we directly provide the entire policy in the context of the LLM and prompt for the binary safe/unsafe judgment.

DPR Implementation Details We use the domain definitions provided in Figure 2 as the policy specification s. Unless otherwise stated, we use Llama 3.3 70B Instruct (Dubey et al., 2024) as $\mathcal{M}$ and Google Search as $\mathcal{G}$. We run research for $k = 3$ iterations and use $n = 20$ for clustering.

A.2 In-House Multimodal Content Moderation

Evaluation We select four domains and provide a one-sentence summary as the input for DPR, as presented in Figure 2. We perform evaluation separately for each domain using the in-domain safe and unsafe examples. Similar to the previous experiment, we provide the entire policy in the context of an off-the-shelf large vision and language model[2] and prompt for the binary safe/unsafe judgment. We experiment in two inference settings: *single sample (S.S)* and *majority vote (M.V.)* with 10 responses sampled with temperature 0.8.

DPR Implementation Details We use Llama 3.3 70B Instruct as the research LLM. We keep all the hyperparameter settings consistent with the previous section and the only difference is the input domain specification.

A.3 Domain Specifications

In Figure 2, we present all the domain definitions for all the experiments. These definitions serve as the seed domain specifications to run DPR.

A.4 Prompts

We provide full prompts for the DPR agent in Figure 3 through Figure 10.

[1]Accessed at https://github.com/openai/moderation-api-release.

[2]We used OpenAI gpt-4o-2024-08-06.

Dataset: OpenAI Online Content Moderation
Policy Domain: Sexual *Definition: Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).*
Policy Domain: Hate *Definition: Content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.*
Policy Domain: Violence *Definition: Content that promotes or glorifies violence or celebrates the suffering or humiliation of others.*
Policy Domain: Harassment *Definition: Content that may be used to torment or annoy individuals in real life, or make harassment more likely to occur.*
Policy Domain: Self-Harm *Definition: Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.*
Dataset: In-House Multi-Modal Advertisement Moderation
Policy Domain: Offensive *Definition: Advertisements that include graphic, gory, vulgar, or culturally insensitive content, or use imagery that shocks or offends viewers.*
Policy Domain: Exploitative *Definition: Advertisements that use shocking, unsafe, or exploitative imagery—especially involving death, disasters, injuries, or sensitive groups.*
Policy Domain: Misrepresentative *Definition: Advertisements that mislead users by implying guaranteed benefits, using false or outdated information, or showing unrelated or unrealistic content.*
Policy Domain: Problematic Finance Claims *Definition: Advertisements that make extreme or misleading financial claims, guarantees, or unrealistic outcomes related to cost, savings, income, or risk.*

Figure 2: Detailed specifications of the domains experimented in this paper. These prompts were created solely for the purposes of this article and are provided for illustrative use only. They do not reflect official Taboola policy, which may be updated or revised over time.

Prompt for Generating Search Queries

```
You are an expert in creating domain-specific knowledge bases. Given a research goal and a
summary of the current knowledge datastore, you write a few querqies to Google for additional
knowledge insufficiently covered by the current knowledge datastore.

Your research goal is: {research_goal}.

The current datastore summary: {current_datastore_summary}.

Write a list of Google queries that would find webpages that expand the coverage of the
datastore. The queries should be in the form of a json list of strings, each string being a
query. The queries should be relevant to the research goal and aim to cover gaps in the
datastore. The queries should be specific. The queries can either directly ask for a specific
information or ask for information from specific source types, which increases the likelihood
of finding the right webpages.

Queries (in json list format):
```

Figure 3: **Prompt for generating web search queries.** The research agent uses it to identify missing coverage.

Prompt for Extracting Rules from a Webpage Chunk

```
You are an expert in creating domain-specific knowledge bases. Given a research goal and
content from Google, you summarize the relevant knowledge in the form of itemized rule.

Based on the following search results generate rules to represent the relevant knowledge.

Generate specific rules that:
1. Are directly extracted or derived from the search results provided.
2. Relevant to the research goal.
3. Cover different characteristics.
4. Are specific. Include any relevant nuances or edge cases mentioned.

VERY IMPORTANT: Your response MUST be a valid JSON array containing objects with these exact
fields:
- "rule": the text of the rule
- "supporting_text": the exact quote from the passage that supports this rule

For example, your response should look exactly like this:
[
{
    "supporting_text": "Direct quote from the passage that supports the first rule",
    "rule": "Rule text goes here"
},
{
    "supporting_text": "Another direct quote that supports the second rule",
    "rule": "Another rule goes here"
}
]

Do not include any explanations, markdown formatting, or additional text before or after the
JSON array.

### Your research goal:
{research_goal}

#### Search Result:
{webpage_chunk}

#### Rule (in json array format):
```

Figure 4: **Prompt for extracting rules from a webpage chunk.** The research agent uses it to generate new candidate rules.

Prompt for Scoring Rule Relevance

```
You are an expert in creating domain-specific knowledge bases. Given a research goal and a
piece of new knowledge you wanted to add to the knowledge datastore, represented as a rule,
judge the relevance of the rule. The rule is relevant if it can be added to the knowledge base
that answers the research goal. If the rule is only broadly related to the research goal,
uninformative to answering the question posed in the research goal, or in the wrong format
(e.g., asking for an action when the research is about definition), it is not relevant. Return
your answer in a json dict with a single key 'relevance' and the value on a scale from 0
(irrelevant) to 10 (perfectly relevant).

##### Research Goal: {research_goal}.


New knowledge (represented as a rule): {rule_text}


##### Is this knowledge relevant enough? Directly write your evaluation in a json dict and do
not write anything else:
```

Figure 5: **Prompt for scoring rule relevance.** The research agent uses it to filter candidate rules.

Prompt for Extracting Rule Keyphrases

```
You are an expert in creating domain-specific knowledge bases. Given the domain description
and an item in the knowledge base, write one keyphrase from the item. The keyphrase should
identify the most salient information (concept or action) that distinguishs the item from the
other items in the knowledge base. The information in domain description itself should not be
in the keyphrase, because it is shared by all the items in the knowledge base.

##### Domain Description: {research_goal}.

##### Item: {rule_text}##### Keyphrase (a single phrase and nothing else):
```

Figure 6: **Prompt for extracting a keyphrase for each rule.** The research agent uses it during datastore indexing.

Prompt for Merging Similar Rules

```
You are an expert in creating domain-specific knowledge bases. Given a domain description and
some items from the knowledge base, combine similar items to make the list more concise. Output
a list of json dicts, each dict corresponding to an item after your processing. Each dict must
have two fields. The first field is "original_items", a list of items you choose to combine,
exactly copied from the original items, and "new_item" is a string for the processed items. You
should not combine items that are dissimilar. For the items you combine, make sure you cover
all the information in the new item but do not write very long sentences. Instead, write a few
shorter sentences to make the semantics clear.

##### Domain description: {research_goal}.

##### Original items:
{rule_text_list}

##### Processed items (output json array of dicts and nothing else):
```

Figure 7: **Prompt for merging similar rules.** The research agent uses it to consolidate extracted rules.

Prompt for Summarizing Section Rules

```
You are an expert in creating domain-specific knowledge bases. Given a domain description and
some similar items that form a single section, generate a short paragraph to summarize the
topic of these items. The summary should serve as a good introduction to this section in the
database. You should take the domain description into account and the summary should
distinguish the items from the other potential sections under the same domain.

##### Domain description: {research_goal}.

##### Section Items:
{rule_text_list}

##### Section Summary (just output the summary text and nothing else):
```

Figure 8: **Prompt for summarizing a section of rules.** The research agent uses it to describe clustered sections.

Prompt for Titling Section Rules

```
You are an expert in creating domain-specific knowledge bases. Given a domain description and
some similar items that form a single section, as well as their associated keyphrases, generate
a title for this section. You should take the domain description into account and the title
should distinguish the items from the other potential sections under the same domain.

##### Domain description: {research_goal}.

##### Section Items:
{rule_text_list}

##### Keyphrases:
{keyphrases_list}

##### Section Title (just output the title text and nothing else):
```

Figure 9: **Prompt for titling a section of rules.** The research agent uses it to label clustered sections.

Prompt for Merging Section Titles

```
You are an expert in creating domain-specific knowledge bases. Given the domain definition a
list of section titles, combine them into a more concise list by merging titles with the same
meaning. Output a list of json dicts, each dict corresponding to an item after your processing.
Each dict must have two fields. The first field is "original_titles", a list of items you
choose to combine, exactly copied from the original items, and "new_title" is a string for the
processed items. You should only combine titles that are similar enough. If the combined title
is so general that it is equivalent to the domain description, do not combine.

##### Domain description: {research_goal}.

##### Existing titles:
{cluster_titles}

##### Combined section titles (in json list format and nothing else):
```

Figure 10: **Prompt for merging section titles.** The research agent uses it to reduce redundant titles.

B Further Analyses

B.1 Analyses on DPR Design

Alternative Research Model. While the main experiments use Llama 3.3 70B for research, we also ran the pipeline with a smaller Qwen2.5 32B (Yang et al., 2024). As shown in Figure 12 and Figure 13, we observe a similar F1 improvement over the baseline, suggesting the generality of the DPR framework.

Convergence Dynamics. Figure 11 tracks the number of unique policy rules and the number of keyphrase clusters throughout five research iterations. The curve exhibits a rapid expansion followed by early stabilisation: over 74% of the final rules are discovered in the first iteration, but only 8% further clusters are added after the second. This plateau indicates diminishing returns beyond $k = 3$ iterations and justifies our choice of the research budget.

Indexing Strategy. Figure 12 and Figure 13 compare three ways of presenting the same policy to the *reader* LLM: (1) a full rule list with the section as the index, (2) a keyphrase *summary* of each cluster, and (3) an ablated "flat" list without section headers. Summaries cut the prompt length but come with a small accuracy drop. The indexed rules and the flattened list have similar performance on Qwen2.5 7B Instruct. However, on Llama 3.1 8B Instruct, the former achieves the best performance. We hypothesize that Llama has a weaker long-context ability and thus can benefit more from more structured input formats.

B.2 Domain Distribution

In Figure 14, we visualize the distribution of domains where DPR draws information from. We observe that DPR does not draw information exclusively from a single domain, nor does it directly copy policy from documents from model providers such as OpenAI that have similar purposes. Instead, it refers to the webpage from a diverse set of domain ranging from Wikipedia, scholarly articles, governmental documents, and various posts on the web. For the OpenAI content moderation, DPR relies more on formal articles to find strict definitions of the domains in interest. By contrast, for the in-house advertisement moderation task, DPR refers significantly more to marketing-related website to find more relevant information.

B.3 Qualitative Study

We provide qualitative studies of webpages, page excepts, and the corresponding DPR-generated rules in Table 3 and Table 4. Overall, we observe that the model performs significant rephrasing and abstraction operations over the raw webpage information. In Figure 15 and Figure 16, we further compare DPR-generated policy with expert-written ones. We find that DPR often can provide rules that are close to expert policy in terms of both style and content. These results strengthen our belief in the potential of using agents to improve the automation in drafting policy documents.

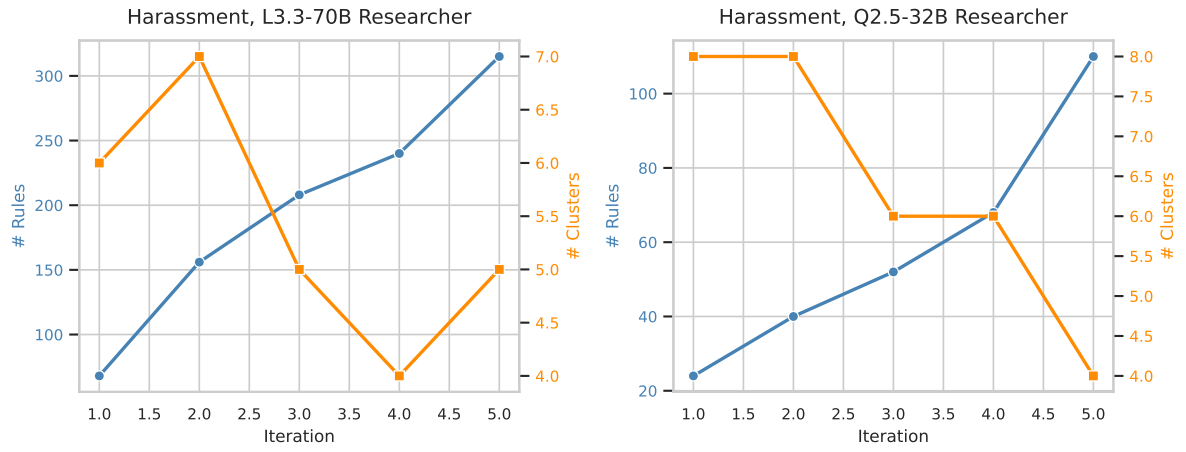

Figure 11: Rule count and cluster count vs. iteration in the Harassment domain for DPR initialized with two LLMs.

Figure 12: Performance vs. iteration in the Harassment domain with two different LLMs for DPR and Qwen2.5 7B Instruct as the reader.

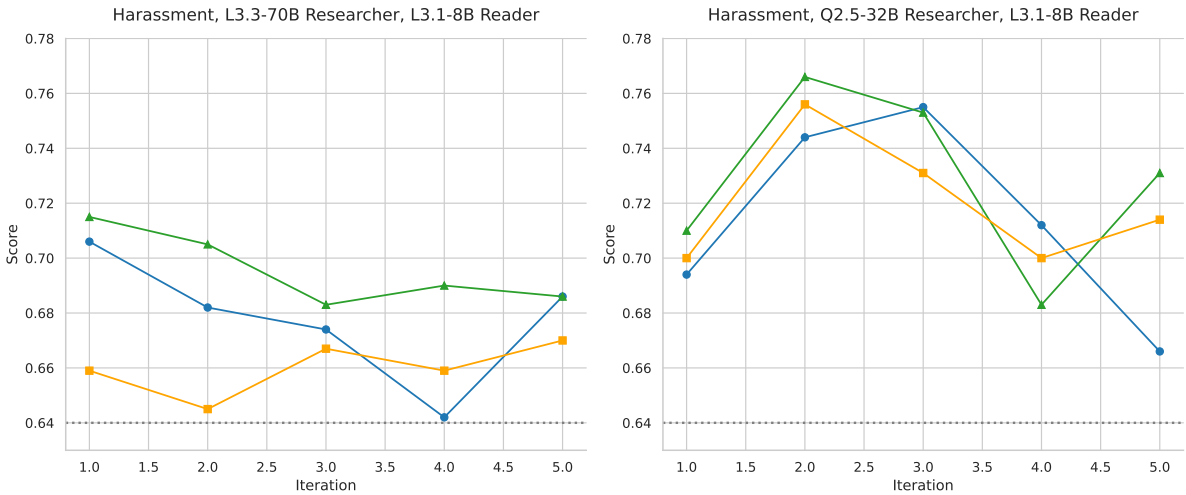

Figure 13: Performance vs. iteration in the Harassment domain with two different LLMs for DPR and Llama 3.1 8B Instruct as the reader.

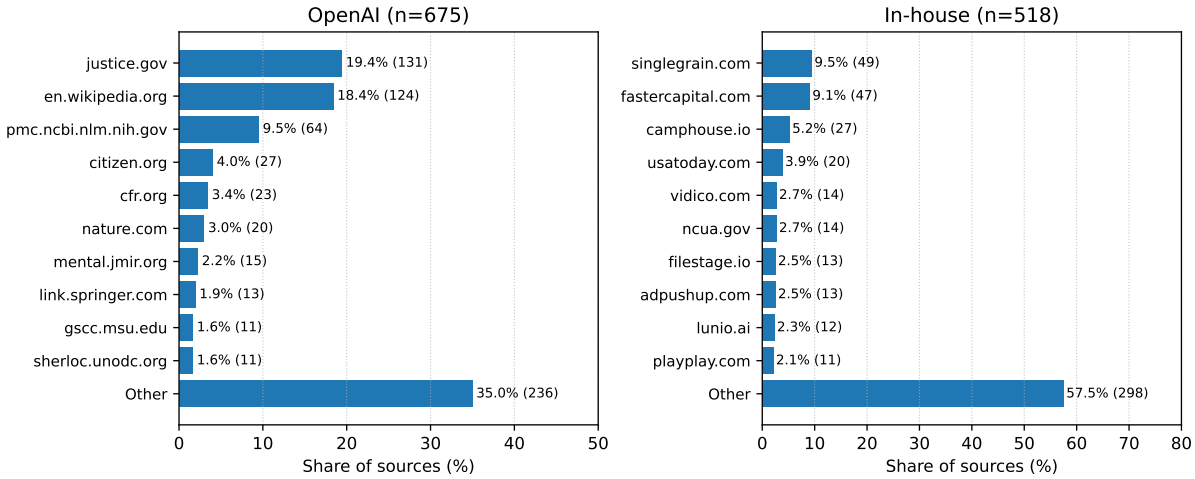

Figure 14: Domain distribution of sources leveraged by DPR for OpenAI and the in-house dataset.

Domain	**URL**	**Page Excerpt**	**DPR-Generated Rule**
Sexual	The Daily Texan (2018)	It found that sex offenders used significantly fewer first-person singular pronouns, such as "I," "me" and "my," and more second-person singular pronouns, such as "you" and "your," compared to the decoys.	Sensitive messages related to sexual content often use fewer first-person singular pronouns and more second-person singular pronouns.
Hate	Wikipedia: Online hate speech	Identity Tourism often leads to stereotyping, discrimination, and cultural appropriation.	Hate messages may involve identity tourism, where a person pretends to be a member of another group. This can lead to stereotyping and discrimination.
Violence	Mastering Cultural Differences	The use of violent language and phrases like "kill two birds with one stone," and "bite the bullet," may seem innocuous on the surface.	Sensitive messages related to violence may contain idiomatic phrases that appear innocuous but contribute to the desensitization of violence.
Harassment	PEN Field Manual Glossary	Dog Whistling: Using words or symbols with a double (or coded) meaning that is abusive or harmful, sometimes to signal a group of online abusers to attack a specific target.	Harassment messages may involve dog whistling, using coded words or symbols with a double meaning to signal other abusers to attack the target.
Self-Harm	Sen. Bennet (Press Release)	In one case, researchers prompted My AI to instruct a child how to cover up a bruise ahead of a visit from Child Protective Services.	Sensitive messages related to self-harm may include instructions on how to conceal physical harm or injuries.

Table 3: Examples of DPR output on the OpenAI content moderation domains.

Domain	URL	Page Excerpt	DPR-Generated Rule
Misrepresentative	SingleGrain: Misleading statistics	If a user finds an anti-aging skincare moisturizer that touts "90% of users experienced fewer wrinkles," this statistic will be enough for consumers to purchase the product. They won't realize or even care that it's a false advertising claim.	Advertisements may use high percentages or statistics to create a misleading narrative, even if the claim is based on biased or incomplete data.
Finance Claims	FMA: Misleading advertising (investments)	What does the advertisement say (or not say) about the risk of the investment? Words such as "secured", "certain", "stable", "guaranteed" can give the impression that an investment is of low or no risk this can be misleading if the investment is not low risk.	Advertisements for financial products should not use words that imply low or no risk, such as "secured", "certain", "stable", or "guaranteed", if the investment is not low risk.
Exploitative	Issuu: Shock advertising examples	PETA's ads often feature graphic imagery of animal cruelty, such as animals being slaughtered or abused in factory farms.	Exploitative advertisements may use graphic imagery of animal cruelty, such as animals being slaughtered or abused, to promote a message or product.
Offensive	Lunio: Controversial ads	Bumble released an ad campaign that sparked outrage online. As shown in the image, the billboard ad campaign included messages such as "A vow of celibacy is not the answer" and "Thou shalt not give up on dating and become a nun".	Using religious references or imagery in a way that is perceived as mocking or disrespectful can be considered offensive.

Table 4: Examples of DPR output on in-house advertisement policy domains.

Human-written Rules and Machine-generated Rules

Policy Domain: Exploitative

Human-written rules.

5.1. Must not use images of catastrophic scenes featuring violence or disaster.
Sub-rules (5.1):
(a) Objects associated with death are permitted (e.g., Coffin, Urn, Tombstone).
(b) Funerals, mourning and scenes associated with death are permitted.
(c) Descriptions of catastrophic scenes are permitted in *Titles* and *Descriptions*.

5.2. Images depicting vehicle accidents must be carefully evaluated for both the visible vehicle damage and the condition of any individuals present. Images that strongly suggest a high-energy impact—with damage and contextual cues indicating a likelihood of life-threatening or severe injuries—must not be used.
Sub-rules (5.2):
(a) Significant vehicle damage is acceptable if individuals exhibit only minor or superficial injuries and the overall context suggests minimal harm (e.g., calm behavior).
(b) Cartoon images are permitted.

5.5. Must not use images of human remains being destroyed or in the process of being destroyed (e.g., coffin entering furnace).
5.6. Must not use images of known personalities who died in the last six months.
5.7. Must not use images of people performing physically dangerous tasks in an unsafe manner.
5.8. Must not use images of marijuana leaves or marijuana plants.
5.9. Must not target a specific ethnicity or religion with dating products or services.

Machine-generated rules (with evidence).

Rule 1. Exploitative advertisements may use images of fatal car crashes to promote their message, even if the circumstances of the crash are unrelated to the advertised issue.
Closest human rule(s): 5.2
Source: Daily Mail article
Excerpt: The sickening campaign shows a picture of a mangled car in a tunnel with the caption: "Diana. She did not choose her death... in 2024, we should have the choice".

Rule 2. Exploitative advertisements may involve the post-mortem exploitation of a dead celebrity's image for financial gain.
Closest human rule(s): 5.6
Source: IJMS article (CCSE)
Excerpt: the post-mortem exploitation of a Deleb's image

Figure 15: Visualization of human-written rules and aligned DPR-generated rules for the **Exploitative** domain. These prompts were created solely for the purposes of this article and are provided for illustrative use only. They do not reflect official Taboola policy, which may be updated or revised over time.

Human-written Rules and Machine-generated Rules

Policy Domain: Offensive

Human-written rules.

6.1. Must not use images of a person with a visible exposed injury (e.g., on skin).
Sub-rules (6.1):
(a) Images of bandaged or cast body parts are permitted.

6.2. Must not explicitly portray in images the inside of a human mouth.
Sub-rules (6.2):
(a) Close ups of forced smiles showing teeth are prohibited.

6.3. Must not use gory or repulsive images.
6.4. Must not use images of a person with a needle or tube inserted into their skin or body.
6.5. Must not use an image of a person with a visible disease, condition or deformity that affects their appearance.
Sub-rules (6.5):
(a) Images of medical procedures are permitted.

6.6. Must not use images of a dead body of a human or animal.
6.7. Must not feature a gun being pointed at the audience or to the edge of the frame.
6.8. Must not feature realistic "fake teeth", realistic veneers, realistic prosthetic teeth or realistic implants without any context.
6.9. Must not feature excessive amounts of realistic blood.
6.10. Must not use images of an unconscious or sedated person.
6.11. Must not use images of destroyed national symbols (e.g., images of burning cash, images of burning flags).
6.12. Must not use offensive language or symbols (e.g., "Take a look at his repulsive home").
Sub-rules (6.12):
(a) Negative language or imagery is permitted.

6.13. Must not use language or images that are culturally or religiously insensitive.
6.14. Must not use vulgar or insensitive language or images.
6.15. Must not feature sensationalized stories about the Amish.
6.16. Must not make the focus of the story a personality as sexual orientation (e.g., "Famous LGBT couples").

Machine-generated rules (with evidence).

Rule 1. An advertisement can be considered *Offensive* if it includes graphic or disturbing content, such as violence against animals.
Closest human rule(s): 6.3
Source: PlayPlay: Controversial commercials
Excerpt: The ad features scenes like firing gerbils from cannons and using a marching band as targets for hungry wolves.

Rule 2. Using religious references or imagery in a way that is perceived as mocking or disrespectful can be considered offensive.
Closest human rule(s): 6.13
Source: Lunio: Controversial ads
Excerpt: Bumble released an ad campaign that sparked outrage online. As shown in the image, the billboard ad campaign included messages such as "A vow of celibacy is not the answer" and "Thou shalt not give up on dating and become a nun".

Rule 3. Offensive advertisements may include lewd or tasteless sexual references, obscenity, vulgarity, brutality, nudity, feces, profanity, or horrifying and repulsive images or words.
Closest human rule(s): 6.14
Source: Wikipedia: Shock advertising
Excerpt: They can include a disregard for tradition, law or practice (e.g., lewd or tasteless sexual references or obscenity), defiance of the social or moral code (e.g., vulgarity, brutality, nudity, feces, or profanity) or the display of images or words that are horrifying, terrifying, or repulsive (e.g., gruesome or revolting scenes, or violence).

Figure 16: Visualization of human-written rules and aligned DPR-generated rules for the **Offensive** domain. These prompts were created solely for the purposes of this article and are provided for illustrative use only. They do not reflect official Taboola policy, which may be updated or revised over time.

C Human Evaluation

We assess the perceived quality and internal applicability of DPR-generated rules by human annotation. The unit of annotation is a single policy rule as it appears in the consolidated, indexed policy for each domain. DPR merges newly generated rules with the previous iteration and organizes them via keyphrase-based clustering into an indexed document. We evaluate the rules in this final index.

Annotation Setup The annotator group comprises three senior content reviewers with domain expertise in policy interpretation and enforcement within the in-house content ecosystem.

Rubric For each domain and each rule, annotators provide the following ratings:

1. Rule clarity & actionability (linguistic):
    - 2 = clear and actionable, easy to apply;
    - 1 = mostly clear but some parts are vague;
    - 0 = too vague to be used in practice.
2. Rule domain relevance:
    - 2 = relevant to domain;
    - 1 = vague (expected to appear in other domains as well);
    - 0 = irrelevant.
3. Internal usability (in-house benchmark only): relevance to the internal use case:
    - 2 = directly relevant (should appear in internal policy);
    - 1 = potentially relevant ;
    - 0 = probably irrelevant.

Results We report counts per label (0/1/2) and mean ± std per metric and domain (Table 5 and Table 6), and inter-annotator agreement using Fleiss' κ with observed agreement $\bar{P}$and chance agreement P_e (Table 7 and Table 8). Across all evaluated domains, the human metrics in Tables 5 and 6 show that most rules are judged clear and actionable and domain-relevant, with the bulk of ratings concentrated at 1–2 on both scales. The small fraction of 0s typically reflects rules whose scope or preconditions require more context than the text provides. On the in-house only internal usability dimension, ratings are likewise skewed toward 1–2 but exhibit more spread, consistent with the fact that organizational constraints (e.g., business model, risk tolerance, and enforcement workflows) can make an otherwise well-formed rule only "potentially relevant." Taken together, the counts and mean±std summaries indicate that the proposed policy synthesis generally yields rules that are understandable, scoped to their intended domain, and plausibly useful for internal moderation, with remaining gaps concentrated in conceptually diffuse or context-heavy areas (Tables 5 and 6).

Inter-annotator agreement (IAA) in Tables 7 and 8 shows moderate to substantial Fleiss' κ overall, with a consistent pattern: κ is typically higher for clarity/actionability than for domain relevance, and lowest for internal usability. This matches with intuition: clarity is more linguistic and thus easier to converge on; relevance requires topical judgment; and internal usability introduces organization-specific priors. Lower κ pockets co-occur with categories that are inherently context dependent (e.g., content that hinges on intent, risk qualifiers, or external claims), signaling where additional guidance would most improve consistency. In short, the IAA results corroborate the descriptive metrics: the rules are broadly serviceable and legible to experts, and the remaining disagreement highlights precisely those areas where policy text benefits from tighter boundaries or richer adjudication cues (Tables 7 and 8).

Domain	Clarity		Relevance		Internal Usability	
	counts	mean±std	counts	mean±std	counts	mean±std
Exploitative	23/36/92	1.4570 ± 0.7460	20/49/83	1.4145 ± 0.7136	20/26/106	1.5658 ± 0.7157
Finance	42/116/331	1.5910 ± 0.6437	43/77/369	1.6667 ± 0.6316	46/81/360	1.6865 ± 1.1256
Misrepresentative	47/47/59	1.0784 ± 0.8314	46/56/51	1.0327 ± 0.7982	53/51/49	0.9739 ± 0.8188
Offensive	46/56/129	1.3593 ± 0.7945	38/78/115	1.3333 ± 0.7441	56/55/120	1.2771 ± 0.8295

Table 5: Human evaluation on in-house test domains: counts of ratings (0/1/2) and mean±std for each metric.

Domain	Clarity		Relevance	
	counts	mean±std	counts	mean±std
Harassment	7/28/172	1.7971 ± 0.4801	9/26/172	1.7874 ± 0.5055
Hate	25/35/231	1.7079 ± 0.6164	27/36/228	1.6907 ± 0.6329
Self-Harm	44/30/51	1.0560 ± 0.8735	37/37/51	1.1120 ± 0.8349
Sexual	14/16/45	1.4133 ± 0.7900	14/20/41	1.3600 ± 0.7822
Violence	17/38/71	1.4286 ± 0.7203	22/47/57	1.2778 ± 0.7445

Table 6: Human evaluation on OpenAI domains: counts of ratings (0/1/2) and mean±std for each metric.

Domain	Clarity			Relevance			Internal Usability		
	κ	$\bar{P}$	P_e	κ	$\bar{P}$	P_e	κ	$\bar{P}$	P_e
Exploitative	0.6069	0.7843	0.4513	0.4146	0.6601	0.4194	0.5802	0.8039	0.5329
Finance	0.0377	0.5399	0.5218	-0.1354	0.5481	0.6020	-0.1280	0.5297	0.5830
Misrepresentative	0.7830	0.8562	0.3374	0.6853	0.7908	0.3355	0.8038	0.8693	0.3337
Offensive	0.7504	0.8528	0.4103	0.5679	0.7359	0.3889	0.8098	0.8831	0.3853

Table 7: Inter-annotator agreement on in-house test domains per metric (Fleiss' κ), observed agreement $\bar{P}$, and chance agreement P_e.

Domain	Clarity			Relevance		
	κ	$\bar{P}$	P_e	κ	$\bar{P}$	P_e
Harassment	0.6337	0.8937	0.7099	0.7021	0.9130	0.7081
Hate	0.6840	0.8900	0.6520	0.7344	0.9038	0.6378
Self-Harm	0.4888	0.6667	0.3480	0.4454	0.6349	0.3417
Sexual	0.7618	0.8667	0.4404	0.5072	0.7067	0.4048
Violence	0.4186	0.6667	0.4267	0.2643	0.5397	0.3743

Table 8: Inter-annotator agreement on OpenAI test domains per metric (Fleiss' κ), observed agreement $\bar{P}$, and chance agreement P_e.