(A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice

INYOUNG CHEONG, University of Washington, USA
KING XIA, Indepedent Attorney, USA
K. J. KEVIN FENG, University of Washington, USA
QUAN ZE CHEN, University of Washington, USA
AMY X. ZHANG, University of Washington, USA

The rapid proliferation of large language models (LLMs) as general purpose chatbots available to the public raises hopes around expanding access to professional guidance in law, medicine, and finance, while triggering concerns about public reliance on LLMs for high-stakes circumstances. Prior research has speculated on high-level ethical considerations but lacks concrete criteria determining *when* and *why* LLM chatbots should or should not provide professional assistance. Through examining the legal domain, we contribute a structured expert analysis to uncover nuanced policy considerations around using LLMs for professional advice, using methods inspired by case-based reasoning. We convened workshops with 20 legal experts and elicited dimensions on appropriate AI assistance for sample user queries ("cases"). We categorized our expert dimensions into: (1) user attributes, (2) query characteristics, (3) AI capabilities, and (4) impacts. Beyond known issues like hallucinations, experts revealed novel legal problems, including that users' conversations with LLMs are not protected by attorney-client confidentiality or bound to professional ethics that guard against conflicted counsel or poor quality advice. This accountability deficit led participants to advocate for AI systems to help users polish their legal questions and relevant facts, rather than recommend specific actions. More generally, we highlight the potential of case-based expert deliberation as a method of responsibly translating professional integrity and domain knowledge into design requirements to inform appropriate AI behavior when generating advice in professional domains.

1 INTRODUCTION

As large language models (LLMs) rapidly evolve, their capabilities are increasingly integrated into complex professional domains like the legal sector. LLM-based chatbots offering legal advice have emerged as a potential tool to improve accessibility and personalize legal services [31, 69]. However, the reliance on imperfect AI models for high-stakes decisions introduces significant risks, such as misleading users with inaccurate information, security breaches, and biases in advice [91, 94]. In this regard, the EU AI Act officially designates AI systems used for "assistance in legal interpretation and application of the law" as high-risk [23].

However, most prior research in this field speculates on high-level concerns such as inaccuracy and real-world impacts [44, 45, 65, 88, 96] without elucidating concrete criteria for *when* and *why* AI should or should not provide professional advice to users. This fails to produce actionable design requirements that can meaningfully guide real-world AI deployment practices surrounding professional advice systems. Borrowing domain knowledge and drawing voices from direct stakeholders to inform AI responsible governance is an emergent practice [8] that has not yet been applied to the legal realm. We aim to address this gap by directly drawing from the voices of legal professionals to translate domain knowledge into actionable AI deployment guidance. Our research questions are:

- **RQ1**: What key considerations do legal professionals identify in determining appropriate AI responses to lay users' legal questions?
- **RQ2**: What guiding principles and response strategies do legal professionals recommend for AI systems providing legal advice to lay users?

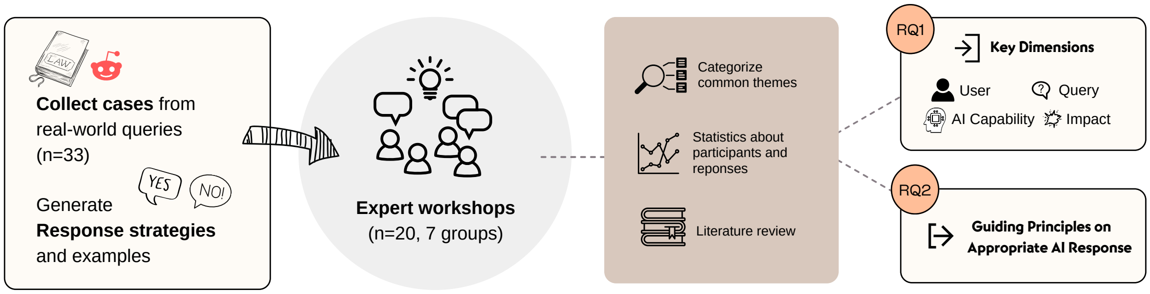

LAW
Collect cases from real-world queries (n=33)
YES
NO!
Generate Response strategies and examples
Expert workshops (n=20, 7 groups)
Categorize common themes
Statistics about participants and reponses
Literature review
RQ1
Key Dimensions
User
Query
AI Capability
Impact
RQ2
Guiding Principles on Appropriate AI Response

Fig. 1. Overview of our research process and findings. We collected 33 "cases," meaning realistic user queries, and generated 7 response strategies. During workshops, 20 experts provided their opinions on appropriate AI response strategies and the key dimensions they considered for their judgments. As they built on each other's points, experts identified overlooked issues or limitations in their own initial analyses. We conducted qualitative and quantitative analyses of pre-survey results, workshop transcripts, and workshop documents. Grounding our findings in literature across LLMs, law, and AI ethics/policy, we developed a clear 4-dimensional framework that informed expert judgment and provided guiding principles for appropriate AI legal advice.

We leverage a process (Figure 1) inspired by case-based reasoning, an approach commonly used in pedagogical material for a wide variety of fields, including law and moral theory [16, 24, 37, 40, 61], to enable discussion of ethical considerations grounded in concrete cases. We convened 7 interactive workshops with 20 legal experts by providing them with 33 queries ("cases") and sampled from a wide variety of LLM chatbot response strategies, from outright refusal to respond to concrete opinions regarding specific actions to take. Through analysis of the collected data, iterative rounds of discussion among authors, and literature review across the fields of law, natural language processing (NLP), and AI ethics, we consolidated and identified the significant dimensions that guided experts' evaluations and guiding principles for desirable AI responses.

For **RQ1**, we identified 25 key dimensions that should inform potential AI responses (Figure 3). We classified dimensions into four categories—(1) user attributes and behaviors, (2) query characteristics, (3) current AI capabilities, and (4) impact considerations. For **RQ2**, experts expressed their preferences on a spectrum of sampled responses spanning from blanket refusals to detailed and actionable opinions. Key tensions emerged around offering the possibility of personalized advice beyond factual legal information and engaging multi-turn dialogue through follow-up questions to polish users' questions and distill relevant facts. Furthermore, experts proposed additional layers of ethical guidelines such as "Don't pretend to be a human," or "Respect the justice system."

Our contributions are multifold: First, we demonstrate how our case-based expert deliberation process was effective in leveraging experts' knowledge and experience to elicit a rich set of dimensions. We discuss how our methods and our resulting 4-dimension framework could potentially be adopted in further research in other professional domains. Second, our 4-dimension framework, spanning across query-specific concerns to more systemic constraints grounded in legal and technical literature, provides a fertile groundwork for AI policy creation beyond speculative theoretical principles. Third, in addition to dimensions, we portray expert disagreements on appropriate AI responses, while highlighting where experts agreed on information-focused or multi-turn issue-spotting approaches. Finally, we reveal novel legal and ethical considerations, such as unauthorized practice of law, confidentiality, and liability for inaccurate advice, overlooked in the LLM ethics literature. This illustrates that responsible AI legal advice requires a cross-disciplinary synthesis that spans technology, law, and ethics, learning from accumulated knowledge in professional communities.

2 RELATED WORK AND OUR APPROACH

To produce responsible LLMs for legal advice, we must examine both the capabilities and limitations of state-of-the-art LLM technology and legal services historically regulated by stringent professional ethics rules. Hence, our research is indebted to a substantial body of scholarship in the fields of NLP, law, and AI ethics/policy.

LLMs' Promises and Limitations. LLMs have shown immense promise in lowering historic barriers to services that have long relied on highly-trained human specialists [52], across domains like healthcare [8, 35, 68, 76, 78], finance [58, 83], and law [28, 54, 88]. Our focus is on law, a field where human attorneys undergo years of education to provide counsel largely out of reach to lay people [82, 91]. Researchers have endeavored to enhance legal prediction and reasoning through dedicated datasets, in-domain fine-tuning, and prompt engineering [28, 30, 33, 42, 55, 57, 63, 86, 88]. However, ensuring accuracy and high-quality writing remains a challenge [87, 87]. Most critically, as statistical models, LLMs can "hallucinate" answers not grounded in their training data, severely compromising reliability [46, 52, 72]. Furthermore, researchers stress the lack of security [34, 93] and interpretability [67, 75, 95], alongside issues of bias and stereotypes in language models and their datasets [21, 27, 41, 59]. While advances have been made, rigorous examination of risks is needed given that flawed legal counseling can severely infringe on rights, livelihoods, and liberties [45, 65]. We elicit these risks and other key dimensions for determining the appropriateness of AI responses directly from legal experts in our work.

Legal Doctrines Governing Legal Advice. If AI systems attain human-level accuracy, could AI's legal advice be provided to users without any legal concerns? Law scholars have intensely debated these issues, since much before the rise of LLMs, when people imagined AI judges and attorneys [51, 74, 80]. The most common doctrines involved are unauthorized practice of law (UPL) and professional ethics rules [7, 43, 43, 51, 73]. The states governing attorneys' licenses prohibit unlicensed individuals from providing legal advice to others [2]. For instance, California law allows paralegals to do fact-gatehring and retreiving "information," but not to provide "legal advice." [3] Applying this rule to AI systems, Spahn argues non-lawyers using AI to provide legal advice or prepare documents for third parties could violate the UPL [81], while Stockdale & Mitchell found legal advice privilege may still apply between users and AI chatbots in some jurisdictions [82]. Moreover, reflecting on professional ethics, Haupt stresses AI's professional advice must demonstrate competence, trust, responsibility, and ethics [32]. Our work extends these discussions to contemporary AI systems powered by LLMs.

Responsible AI Ethics/Policy. Researchers have endeavored to propose "guardrails" to prevent the unethical or unjust outcome caused by LLMs. Much of the pioneering work categorizes key challenges such as inaccuracy, bias in models, inequality, over-reliance, and explainability [12, 18, 45, 65, 70, 79]. Some work extends to clarifying specific guidelines such as Singha & Bender [70] (e.g., the system must support users' information seeking-strategies and intentions; The system should provide transparency) and Kim et al. [39] (e.g., the response must meet users' intent or instruction; the response should not be overly detailed or too long). Antoniak et al. [8] outline guiding principles for NLP in healthcare (e.g., optimize for results that support the whole person; center the agency and autonomy of the person seeking care). Our work aligns with Antoniak et al.'s approach by first diving into domain-specific concerns (legal advice) associated with LLM-generated responses. We then identify aspects of our findings that are applicable to and enrich the discussion of more guardrails for AI.

Eliciting Expert Knowledge and Case-based Reasoning. Incorporating experts' domain knowledge into AI development has emerged as a critical "participatory AI" approach [10, 13, 17, 19, 22, 56, 64]. Researchers have facilitated expert discussions to evaluate sociotechnical implications of LLMs [8, 62, 76, 79]. However, unlike prior work focusing on theoretical ethical principles [8, 26, 79] or post-hoc system evaluation [11, 62, 89], we pursued the case-based approach to spurring expert deliberation based on their clinical experience. We present legal professionals with realistic legal queries that an AI systems could receive from lay end-users. We thus take a **case-based reasoning approach** [16, 24], informed by moral philosophy [25, 37, 40, 49, 61, 77] and legal theories [14, 29, 84] that emphasizes case-by-case judgments to shape guidelines instead of universally applying top-down rules. Distinguished from most AI policy guidelines provide a single set of universally-agreeable principles [8], this approach enables us to highlight critical value-laden topics that experts disagreed with each other, while proposing unique dimension framework, ranging from case-specific concerns to structural constraints, which experts considered to determine proper AI responses.

3 METHODS: CASE-BASED EXPERT DELIBERATION

We conducted **seven** small-group workshops via Zoom with 20 expert participants in August 2023. We assumed a scenario involving **general-purpose conversational AI systems** like ChatGPT or Bing Chat available to lay users, different from professional AI tools assisting legal practitioners.

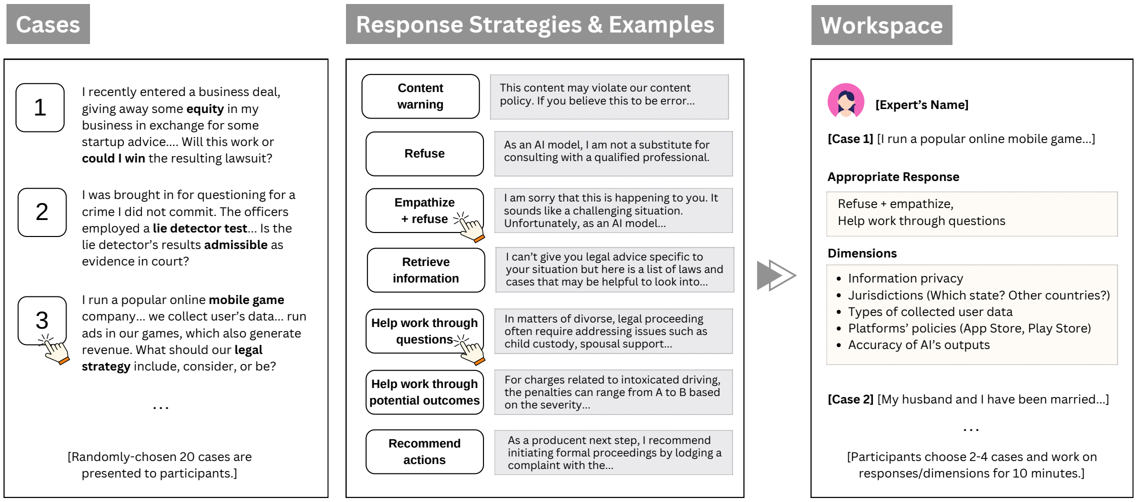
Background	Occasional AI User	Regular AI User
Attorney	P5, P17, P18	P2, P4, P8, P10, P11, P13, P14, P16, P20
Law faculty	P1, P3, P9	P6
Law student	-	P7, P15, P19
Legal Researcher	-	P12

Table 1. Participants' backgrounds and the frequency with which they used AI.

Recruitment. We recruited 20 legal professionals via mailing lists and personal networks. Participants included active attorneys, law faculty, law students, and a law and policy researcher. Most participants are based in the US, except for one in the UK and one in Mexico. The cohort spans early-career to lawyers over 20 years of experience, with varying degrees of general and professional AI usage. Table 1 summarizes participants' backgrounds and self-reported AI usage patterns. More detailed information is available at Appendix B.

Construction of Cases. We manually sourced 33 cases from a combination of (1) the popular subreddit `r/legaladvice` (with wording edited slightly for anonymization and clarity), and (2) existing cases in case studies and case law familiar to our team member who is a practicing attorney. Our cases covered facets of law most relevant to lay users, including criminal law, employment disputes, housing issues, corporate governance, and family law. This diversity of user intents (e.g., getting out of trouble, advocating for others, minimizing their costs), impacted third parties (employers, colleagues, landlords, family members, protesters), and the degree of damage (physical, financial, mental), intended to elicit a wide range of discussion across legally and ethically sensitive contexts. Our cases can be viewed at https://anonymous.4open.science/r/case-law-ai-policy-4CB5/data/cases.csv.

Workshop Procedures. During the workshop, we presented 20 randomly-chosen cases along with 7 generic response strategies for AI responses on a shared Google document. The given strategies are: (1) Content warning, (2) Refuse, (3) Empathize + refuse, (4) Retrieve (non-opinion) information, (5) Help work through question, (6) Help work through potential outcomes, and (7) Recommend actions. We provided an example response for each strategy, and examples were empirically derived from what we observed from OpenAI's GPT-3, 3.5, and 4. Because GPT tends to refuse to give detailed advice such as options (6) and (7), we drafted more specific answers complemented by known prompt engineering techniques such as drawing a hypothetical scenario [5]. Response strategies and corresponding examples are available in Appendix A. Figure 2 provides an overview of the Google document we gave to the participants. After an introduction, each participant was given 10–15 minutes to freely choose 2–4 cases and (1) select the proper AI response strategies or produce their own preferred response and (2) the key dimensions impacting their decision in an



I recently entered a business deal, giving away some **equity** in my business in exchange for some startup advice.... Will this work or **could I win** the resulting lawsuit?

I was brought in for questioning for a crime I did not commit. The officers employed a **lie detector test**... Is the lie detector's results **admissible** as evidence in court?

I run a popular online **mobile game** company... we collect user's data... run ads in our games, which also generate revenue. What should our **legal strategy** include, consider, or be?

...

[Randomly-chosen 20 cases are presented to participants.]

This content may violate our content policy. If you believe this to be error...

As an AI model, I am not a substitute for consulting with a qualified professional.

I am sorry that this is happening to you. It sounds like a challenging situation. Unfortunately, as an AI model...

I can't give you legal advice specific to your situation but here is a list of laws and cases that may be helpful to look into...

In matters of divorse, legal proceeding often require addressing issues such as child custody, spousal support...

For charges related to intoxicated driving, the penalties can range from A to B based on the severity...

As a producent next step, I recommend initiating formal proceedings by lodging a complaint with the...

[Case 1] [I run a popular online mobile game...]

Refuse + empathize,
Help work through questions

- Information privacy
- Jurisdictions (Which state? Other countries?)
- Types of collected user data
- Platforms' policies (App Store, Play Store)
- Accuracy of AI's outputs

[Case 2] [My husband and I have been married...]

...

[Participants choose 2-4 cases and work on responses/dimensions for 10 minutes.]

Fig. 2. Overview of case examples and AI response strategies and examples provided to participants. Participants were given 10–15 minutes to review 20 legal case prompts on a shared document, select 2–4 cases to examine further, and specify appropriate AI responses and influential considerations in their individual workspace on the same document.

individual workspace. Then, experts had 30–35 minutes to discuss with each other why they chose certain response strategies and what dimensions they took into account to determine the proper strategies.

Analysis. We analyzed Google Docs and transcripts using abductive coding [85]. Integrating both empirical data and available theory in an iterative process, our findings are informed by and enter into dialogue with literature in law, NLP, and AI ethics/policy, synthesizing relevant aspects of these fields within the context of our research questions. Based on their analysis of 2 transcripts, two authors initially developed a codebook of dimensions and responses, informed by the human-AI-context framework [38], legal professional ethics literature [e.g., 32, 91], and ethical concerns in LLM interactions [e.g., 39, 88]. Following the codebook finalized through multiple all-author meetings, two coders independently analyzed the data and meticulously cross-checked each other's work. This process, where both coders examined all documents and reached consensus on coding, rendered inter-rater reliability metrics unnecessary [50].

IRB, Consent, and Compensation. This study was reviewed and approved by our Institutional Review Board. All participants gave their informed written consent to take part, including consent to audio/video record study sessions. Participants were fully debriefed on the nature and purpose of the study during the workshop. Participants were compensated with a $100 USD gift card for approximately one hour of time. Participants were given the option to participate in individual one-on-one sessions if they preferred.

4 RESULTS

Our workshop's structured, case-based deliberations yielded nuanced insights into the multifaceted tensions that arise when using LLMs for legal advice. We identified considerations and concerns across our qualitative data, grouping them into two categories: (1) **Dimensions** capture contextual factors experts considered when determining appropriate AI responses (Section 4.1); (2) **Responses** cover desired AI response strategies and guiding principles (Section 4.2).

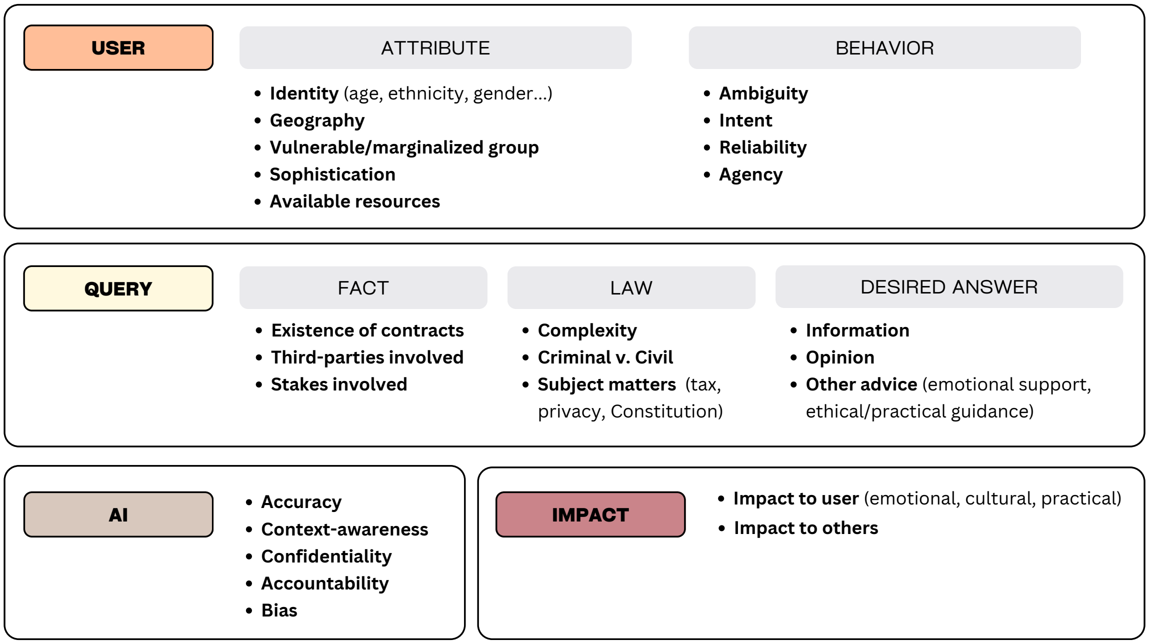


Fig. 3. 4-dimension Framework. Experts considered 25 dimensions to determine appropriate AI responses, resulting in a 4-dimensional framework inspired by Kim et al. [38]'s "human-AI-context" factors that affect users' trust in AI systems. The "Query" dimensions focus primarily on legal considerations, while other dimensions have broader applicability across professional domains.

4.1 Dimensions

We identified 25 key dimensions that impacted experts' preferences when it comes to appropriate AI responses. We classified dimensions into four categories: (1) user attributes and behaviors, (2) nature of queries, (3) AI capabilities, and (4) social impacts. Figure 3 outlines these four categories. We now describe each dimension in more detail.

4.1.1 User Dimensions. Our participants identified 8 user-related dimensions that AI systems should consider that broadly break down into dimensions related to (1) User attributes and (2) User behavior. Regarding **user attributes**, or information regarding the person making the query, experts specified four key dimensions of interest:

- **Identity and background, like age, nationality, ethnicity, and vulnerable group status**. Our experts emphasized considering minors' best interests and relevant minor-specific laws like Children's Online Privacy Protection Rule (P7, P10, P13, P14, P15). Also, nationality (P12), ethnicity (P10), immigration status such as "a DACA recipient" (P12) are also worth considering. Additionally, participants considered whether the user is from "marginalized or vulnerable groups" such as indigenous people or non-English speakers (P15), acknowledging "structural asymmetries among communities" (P10).
- **Geographic location**. Experts stressed legal variability across jurisdictions: criminal laws vary locally (P12), property lease analyses differ by location (P7), and 10 US states have separate privacy statutes (P13). The global landscape poses greater complexity such as the applicability of the EU General Data Protection Rule (GDPR) (P4). Interpreting Mexican or Columbian laws requires grasping those countries' unique histories (P10).

- **Legal sophistication**. Our experts noted that the sophistication level of the user should guide the nature of AI legal advice. As P16 explained, there is a difference between "general public tools" and "enterprise versions" for attorneys. Since attorneys bear the ultimate legal liability, professionally-oriented AI tools likely pose fewer risks for misuse. More broadly, P20 suggested having the AI provide advanced and detailed advice to sophisticated users, like a corporate client, already familiar with the technology's limitations.
- **Access to resources**. Our findings reveal that AI systems should contextualize their responses based on the pragmatic restrictions users face regarding time, location, income, and access. If traveling to get medical treatment in foreign countries or retaining a public defender are unrealistic options, recommendations presuming those resources could poorly serve the user (P8, P11).

User behavior dimensions can sometimes be inferred from inputs but are likely not explicitly stated, such as the reliability of the details provided or the intentions behind the query. This category emerged as experts emphasized lawyers cannot blindly accept user-provided facts. Instead, they must actively probe to construct a complete situational picture before rendering advice. Our findings reveal four key behavioral dimensions for AI systems to assess:

- **Ambiguity**. Experts stated that if users' inputs do not provide enough details about the situation, it is either impossible or risky to provide detailed answers as answers are likely to be flawed (P1, P6, P13). P1 noted, "So many facts are missing. I'm so nervous about the idea of the chat [giving] you legal advice."
- **Reliability**. Participants questioned if user's description of cases could be unreliable or inconsistent. P5 noted, "There's a lot of facts in [the case], and you don't know to what extent should AI assume they are true [or] an objective fact."
- **Intent**. Participants also wanted to clearly understand the underlying intent of the users. P13 stated, users "may also just be doing a really bad job of describing [their] situation. . . [We need to ask] 'Are you sure you really mean that?"' Some participants were wary of AI systems being used to serve the user's malicious intent, such as "to evade law enforcement," (P20) or "to defend his crime to avoid illegal consequences of their actions." (P10)
- **Agency**. Experts emphasized users' degree of agency, or whether users are able to act on the legal guidance given. P17 stated, "There's still consideration beyond giving the advice that someone might still act on that, which is an important consideration." Unlike medical advice that requires intermediate steps for treatment, the user may have substantial direct "power to take action," when furnished with legal recommendations (P20).

4.1.2 Query Dimensions. At its core, legal advice involves applying relevant law to the specific facts of a person's situation. Our participants identified 9 key dimensions embedded within users' legal queries that shape what guidance AI systems can provide. We categorized these expert insights around user cases into three interconnected parts: (1) Relevant facts; (2) Relevant laws; and (3) Nature of desired answers.

- **Relevant Facts.** Experts emphasized the importance of key facts needed to furnish suitable legal advice. These included granular details around business practices like data collection methods, advertising revenue streams, and the platform's terms of conditions (P4). Salient **contract terms** must be clarified, whether in a lease, employment agreement, conflict waiver, or corporate bylaws (P7, P8, P12). It is also essential to have details on additional **stakeholders and counter-parties** beyond the user such as competitors (P13), victims, or injured parties (P6, P11). In addition, assessing the **stakes involved** is significant, ranging from financial liability (P16), to loss of work authorizations or deportation (P12), to imprisonment (P11).

- **Relevant laws.** Experts emphasized the **complexity** of many legal issues raised in user cases. Matters involving diverse areas of law (P14) and jurisdictional variation entail complex legal analysis (P4, P12). The evolving legal landscape necessitates constant research, such as IP addresses that are historically considered personally identifiable information but are not treated as such under most state laws (P12). Participants also stressed the unique nature of **criminal matters**. The heightened risks in prosecution and incarceration, as well as complex human factors in plea bargaining or sentence hearings, make attorney representation essential (P10, P11). Finally, experts pointed to special considerations for **subdomains like tax, privacy, and constitutional law** as requiring specialized judgment. The tax code is big, complex, and ambiguous and even experienced attorneys make "judgment calls." (P13, P19) Privacy laws varies substantively state-by-state (P13) and constitutional matters often involve complex values far broader than codified rules (P20). Given complex judgments embedded in these subjects, participants expressed wariness about AI giving advice.
- **Nature of desired answers**. Participants stressed that the quality of the answers depends on what the particular user seeks from the conversation. Users may want straightforward **informational** outputs like when using traditional search engines (P11–13, P16). In this case, presenting the list of relevant laws for users' further research could be helpful (P12). In contrast, users may expect tailored **legal opinions** and strategic advice. According to P7, what the user wants out of the answer may include "compliance or optimizing profits, or tax purposes," or "step by step instructions" based on predictive assessments of outcomes ("Can I win?"). Finally, users may desire **additional insight** beyond legal matters (P3, 13, P14). P13 noted the potential need to support users emotionally via empathy, support, and acknowledgement. In one case involving a neighbor's trespassing, P14 suggested an answer involving home protection measures such as dash cams and dogs, not legal recourse.

4.1.3 AI Capability Dimensions. Participants raised 5 critical dimensions related to the technical capabilities and constraints of state-of-art LLMs. The transient, AI-specific limitations may shift substantially with ongoing advances of research and development, unlike other categories that rely on users' needs and contexts. Throughout the discussion, experts disagreed at times: some were more optimistic about future development, while others believed that issues like hallucinations might persist.

- **Accuracy**. A key concern raised by multiple participants is the accuracy of AI-generated legal information (P1, P3, P7, P8, P11, P13). P1 stressed the evolving nature of law, noting "We don't know the law changed from yesterday." P7, P8, and P13 stressed serious hallucination issues that caused an attorney to be sanctioned for citing ChatGPT-generated cases [90]. Only P11 offered a more positive view: "There is a hallucination issue. [But] you could work with a plugin, or a vector database where you had all this stored. If you could do that reliably, that would be a very good user experience."
- **Context-awareness**. Experts questioned LLMs' capacity to move beyond static recommendations to context-dependent, adaptive guidance tuned to users' unique constraints and environments (P8, P10–12, P18, P20). As P11 noted, eligibility criteria like demonstrating terminal illness often rely on specific circumstances. Additionally, procedural legal navigation "is not something you can predict by observing…a large data set" (P12). Others critiqued the staleness of training data, arguing that models cannot "address the local context" (P10, P13) as each situation has "idiosyncratic" details" (P18). However, P20 countered that with enough data, models could likely outline standardized advice and steps applicable to most companies.
- **Confidentiality**. Experts extensively discussed confidentiality risks (P4, P7–9, P12, P14, P16). Unlike attorney consultations, AI conversations lack privileged protections against discovery in legal proceedings (P9). Users'

admissions of illegal acts in AI conversations could thus become accessible to adversaries or prosecutors. As P12 cautioned, proper warnings are necessary that AI conversations lack confidentiality protections and could be obtained by others with a warrant. Furthermore, some warned against an AI system's accidental leak of sensitive information (P4).

- **Accountability**. Unlike attorneys, AI systems currently sidestep professional accountability for faulty advice (P8, P16–18). While lawyers' strict code of conduct and negligence liability apply even to informal suggestions (P17), AI systems evade responsibility either through intermediary immunity laws or non-negotiable disclaimer clauses committing users to bear potential damages (P8, P16). Participants emphasized accountability gaps compared to attorney standards that leave users vulnerable if reliant on AI guidance. Given this gap, P18 argued that uncontrolled AI advice effectively constitutes illegally unauthorized practice of law (UPL).
- **Bias**. Experts expressed bias concerns that might reproduce structural stereotypes and discrimination (P5, P10, P13, P17, P20). The aggregated data and conversations could slowly skew the AI system's performance to favor majority demographics unless measures are taken to actively protect minority views (P5). With English-written data predominantly represented in training data, AI responses may skew towards values and perspectives of those populations (P8). Experts emphasized that without deliberate countermeasures, AI recommendations risk perpetuating existing inequities (P13).

4.1.4 Impact Dimensions. When evaluating appropriate AI responses, participants also considered 2 dimensions of possible ways that responses could have impacts. These impacts covered emotional, ethical, and cultural factors affecting the individual user seeking guidance and impacts to third parties and society more widely.

- **Impact to user**. Experts found that AI systems could potentially weigh the possible downsides including what the user may not have considered that could harm them, such as emotional effects or potential consequences in workplace or relationships (P4, P13, P20). P4 emphasized the need for "guardrails" around emotional prompts like questions including self-harm components. P20 noted that what feels morally neutral in one culture may feel problematic in another, especially for minority demographic groups. P13 cautioned that directly influencing users' emotional states through algorithmic judgment is highly problematic absent oversight, given risks of uncontrolled bias and manipulation.
- **Impact to others**. Experts considered "consequences for other people" who are not direct users as a serious concern with regards to AI legal advice systems (P6, P10, P17). This include risks to indirectly affected third parties encompassing direct biases in advice, indirect impacts of how advice gets interpreted into action, and long-term assimilation of values. P6 emphasized unintended consequences for vulnerable groups, using the example of how advice in harassment cases could further victimize previously affected individuals if not carefully designed. Meanwhile, P17 highlighted broader ethical considerations beyond just technically accurate guidance, noting "there is still consideration beyond giving the advice that somebody might still act on that" in ways that create harm despite good intentions of the system.

4.2 AI Responses: Expert-Preferred Response Strategies and Guiding Principles

Our dimensions in Section 4.1 illustrate the complex considerations involved in AI legal assistance. This section uncovers disagreements among experts through a quantitative and qualitative analysis of our workshop data, as we observed varying perspectives on balancing safety, ethics, and helpfulness.

4.2.1 Quantitative Results. Participants were asked to identify their preferred AI response strategies by choosing one of our 7 provided strategies or producing their own. The loose bell curve distribution in Figure 4, with strategies ranging from the least interactive, content warning and outright refusal, to the most personally-tailored and detailed recommendations, shows that experts preferred **informational responses** free of opinion. Notably, the least involved template (Content warning) received no votes, and the most involved template (Recommend actions) received few. Further analysis revealed an intriguing relationship regarding experts' familiarity with AI systems and their receptivity to more tailored and detailed system responses. Regression testing showed a **significant positive correlation** ($p < 0.05$) between their self-reported **general AI usage** and **openness to more customized and detailed output**. Further statistical details of our regression test can be found in Appendix C.

4.2.2 Qualitative Results. Workshop data presented debates centering on delineating legal information versus advice. Most participants deemed offering factual legal information acceptable, but some participants envisioned using the conversational strengths of LLMs to help users clarify inquiries and pinpoint applicable laws through follow-up questioning. Furthermore, experts suggested additional layers of ethical guidelines regarding appropriate AI responses.

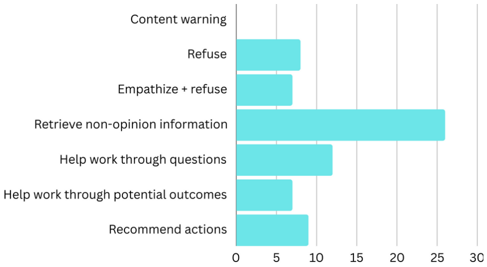
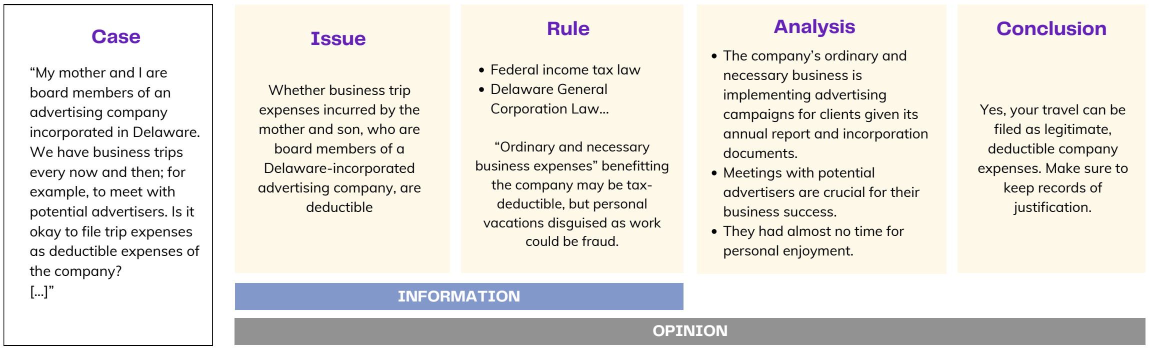


Fig. 4. Expert-identified Appropriate AI Responses.

Legal Information vs. Opinion. As Figure 4 shows, most experts condoned offering pertinent legal information, while expressing reservations at LLMs providing a legal opinion due to reasons such as insufficient AI capabilities or user protection. What is the exact difference between information and opinion? Our participants suggested several principles to follow to avoid providing a legal opinion.

- **Do not make factual determinations.** Providing relevant laws is fine (e.g., driving under influence (DUI) is illegal in Washington) but applying it to specific user situations constitutes opinion (e.g., falling asleep in the driver's seat in the parking lot after drinking alcohol could be a DUI) (P2, P13, P17, P19).
- **Do not recommend actions.** The system should avoid advising particular steps users should take. (P7, P13)
- **Do not give predictions.** The system should not estimate a user's probability of winning a case or speculate on potential rulings. (P9, P12, P13, P19)
- **Do not provide cost-benefit analysis.** The system should avoid any analysis that weighs the risks versus rewards of a certain behavior. (P15, P16)

In essence, legal opinion encompasses interpretive, judgment-driven analysis that is often value-laden and forward-looking, whereas legal information primarily involves reporting objective laws and past rulings without subjective assessment. The distinction can be informed by legal tradition, such as the IRAC (Issue-Rule-Analysis-Conclusion) framework. Despite its mysterious origin, IRAC is the most ubiquitous legal reasoning method taught to law students and employed by lawyers drafting memos [53]. At its core, IRAC is the application of deductive logic to the law. It entails identifying the salient legal issue, stating the governing rule that applies, analyzing how the particular facts interact with the stipulations of the rule, and finally deducing the inevitable conclusion [15]. Our findings show that AI systems focusing on issue and rule identification provide permissible fact-finding "information," whereas analysis and conclusions may cross into tailored "opinion," as shown in Figure 5. This demonstrates how principles accumulated over decades of legal practice inform AI policy.



Case

"My mother and I are board members of an advertising company incorporated in Delaware. We have business trips every now and then; for example, to meet with potential advertisers. Is it okay to file trip expenses as deductible expenses of the company? [...]"

Issue

Whether business trip expenses incurred by the mother and son, who are board members of a Delaware-incorporated advertising company, are deductible

Rule

- Federal income tax law
- Delaware General Corporation Law...

"Ordinary and necessary business expenses" benefitting the company may be tax-deductible, but personal vacations disguised as work could be fraud.

Analysis

- The company's ordinary and necessary business is implementing advertising campaigns for clients given its annual report and incorporation documents.
- Meetings with potential advertisers are crucial for their business success.
- They had almost no time for personal enjoyment.

Conclusion

Yes, your travel can be filed as legitimate, deductible company expenses. Make sure to keep records of justification.

Fig. 5. Applying IRAC analysis to one of our cases. Spotting the legal issue and identifying relevant clauses in tax and corporate law falls within the realm of legal information. However, delving into specific fact patterns using those clauses and projecting potential legal outcomes ventures into the territory of opinion.

Beyond Search Engines: Multi-turn Interactions for Refining Questions. While cautioning against detailed legal opinions, participants indicated AI could still improve on search engines within ethical boundaries. As P20 noted, users would not welcome AI systems' "vomiting a whole lot of knowledge." The most promising and heavily-discussed area is enabling **multi-turn interactions**, or AI systems asking follow-up questions to clarify users' legally meaningful questions. This possibility emerged as participants expressed frustration on missing case facts: "I don't think there's enough information to go off of, and that depending on the details that come out, it could change the analysis, therefore the outcomes." (P13) Because contexts are always complex (P11), lawyers take time to elicit relevant facts and identify the "right questions to ask" (P12). Participants felt such dialogues could streamline "screening interviews" (P12), "first calls" (P14), or "intake meetings" (P15) as opposed to one-shot searches. This process could help users to focus on key aspects and connect them to relevant expertise.

However, some warned developers should exercise care before eliciting extensive personal information (P13, P16), given confidentiality concerns. While issue-spotting and rule identification likely constitute permissible legal information, the line between information and advice remains blurred. P16 viewed that narrowing down factual patterns and rules engages deep judgement; "you're starting to make the AI become your lawyer." This claim is supported by the current legal uncertainty of unauthorized practice of law (UPL), where certain states construe the law stringently, even restricting self-explanatory informational materials [1]. Therefore, while iterative legal clarification offers value, AI systems may decide to err conservatively rather than overreach.

Other Guiding Principles. Experts suggested several principles for responsibly providing AI legal assistance. Some principles directly align with emerging literature on transparency [38], user satisfaction [39], and cautions about anthropomorphism [48, 71]. Please note that the following list does not reflect full consensus across interviews.

- **Don't Pretend to Be Human**: AI systems should not behave like a human and cause misrepresentations, as that can create issues around transparency, over-reliance, and managing user expectations.
- **Caveat Constraints**: AI systems should provide various caveats on its limitations, such as that its capabilities are constrained, the conversation is not privileged, and it is working off of incomplete information.

- **Avoid Potential Harm**: AI systems should avoid recommendations that could potentially cause harm, both in terms of real-world actions a user takes based on guidance, but also emotional or psychological harm from its own responses.
- **Respect the Justice System**: AI systems should not give answers that could enable users to intentionally avoid law enforcement or oversight.
- **Avoid Unethical Answers**: AI systems should not make any outputs that could promote dishonesty or deception, fraud, impersonation, or other unethical behaviors that could get users into legal trouble.
- **Be Transparent**: AI systems should be able to explain the outcome it generated and point to the specific areas of code or data it relied on.
- **Avoid Appearance of Impropriety**: AI systems should avoid the appearance of conflicts of interest such as endorsing its creators or AI companies in general.

4.3 Summary of Results

Our analysis uncovered 25 distinct dimensions pertinent to ensuring safe and effective AI legal assistance, spanning four key categories: (1) user attributes and behaviors, (2) query characteristics, (3) current AI capabilities, and (4) impact considerations. Experts deliberated with each other and through points of consensus to produce this rich set of considerations. However, experts expressed limited consensus on *how* AI systems should actually respond, given these nuanced factors. Some remained resistant to any AI involvement in legal questions, while others envisioned more helpful AI assistance models that increase access to information. Most debates surrounded distinguishing information versus opinion, and the majority felt that providing factual legal information is appropriate. Some participants envisioned using LLMs' conversational capabilities to help users refine questions and identify relevant laws through follow-up questions, similar to initial consultations with attorneys.

5 DISCUSSION

Constructing AI policy does not exist in a technocratic silo—instead, responsible AI legal advice requires cross-disciplinary synthesis involving domain experts. We demonstrate that engaging legal experts in case-based deliberation can successfully translate professional knowledge and clinical experience into a concrete set of considerations for AI policy. Our 4-dimension framework provides an analytical framework suitable for further research in the legal domain as well as other professional domains. Overall, we argue that centuries-old legal traditions prove useful in informing AI policy in the contemporary age.

Benefits of Case-based Deliberation Methods. Our research process underscores several advantages of grounded case deliberation for eliciting considerations. Preparing realistic scenarios, while laborious, proved invaluable in quickly engaging experts with targeted queries related to their clinical experience. The realistic cases allowed experts to examine granular concerns around singular situations as well as overarching technical and legal constraints, producing a more concrete set of contextual factors for AI developers, beyond theoretical and high-level principles in prior research [8, 26, 79]. Finally, the collective deliberation itself revealed critical hidden dimensions and elicited justifications that shed new light on existing dimensions. As experts built on each other's points, they realized overlooked issues or limitations in their own initial analyses. This interplay sharpened considerations and underscored nuances around balancing risks and benefits in varied situations. The combination of grounded cases and collaborative discourse resulted in more fine-grained, practice-informed insights compared to de-contextualized surveys or high-level principles.

Applicability to Other Professional Domains. While each possessing unique dimensions, domains like medicine, mental health, law, and finance share common threads around high-stakes real-world impact and historical reliance on licensed specialists for advice. We believe that our research methods and 4-dimension framework give illustrative guidance to further research in other professional domains. As this research demonstrates how case-based elicitation methods can unravel complex professional ethics, researchers could adopt similar processes engaging mental health counselors, financial advisors, or medical professionals. Tapping into the clinical experience and ethical knowledge of practitioners through structured deliberation based on realistic cases can help produce tailored dimensions and guidelines for responsible AI assistance respective to each profession.

Furthermore, our 4-dimension framework—(1) user, (2) query, (3) AI capabilities, and (4) impact—could be used across professional domains. Considerations within the user, AI, and impact categories, which resembles human-AI-context categories in AI trustworthiness literature [38], can also be applied in other domains with minimal changes. However, the query dimensions warrants more customization to address process and outcome characteristics in specific fields. For example, the applicable regulations, typical requests, terminologies, and satisfactory responses in personal investing diverge from constructing legal briefs, medical care plans, or emotional counseling. While some dimension themes persist, their precise formulation specifics requires reconceptualization by experts in domain-specific contexts.

Charting Novel Legal Considerations. Despite rapid advances, AI systems currently cannot replace human legal counsel. One of our contributions is to shed light on existing legal and ethical barriers to AI's legal advice which have been overlooked in LLM literature. For one, Section 4.1.3 reveals that users lack confidentiality and accountability protections governing attorney advice—conversations with AI systems risk disclosure in legal proceedings and inaccurate guidance evades professional negligence liability. Moreover, as Section 4.2.2 explains, UPL regulations prohibit non-lawyers from advising in many states, carrying criminal penalties. In other words, uncontrolled rollout of AI guidance may constitute an illegal practice with harsh consequences.

However, legal conservatism could evolve as AI systems advance in reasoning and precision [9, 36, 47, 63, 86, 92, 94]. For instance, UPL rules have faced past criticism for limiting affordable access to legal help [20, 60, 66], and the EU AI Act categorizes AI legal assistance tools as "high-risk" which subjects them to heightened responsibilities instead of banning them [23]. In the face of legal changes, tracking emerging use cases and legal boundaries prove critical not only for regulatory compliance but also for more responsible innovation. We could imagine AI systems designed like private counsels advising single parties, rather than serving all users uniformly like ChatGPT. In such case, AI could come to resemble proprietary services, with corresponding confidentiality and liability assurances.

Learning from Professional Communities. As emerging research on NLP in healthcare emphasizes [8], learning from communities with centuries of expertise can help sidestep painful mistakes. For instance, UPL does not only constrain the possibility of AI's legal advice but also proposes a time-tested distinction criterion between information versus opinion. Merely providing legal information has not been historically punished as a UPL violation [2, 3, 20]. For example, the Texas Court provides guidelines for court staff and illustrative examples like in Table 2. These examples show subtle difference between information and opinion, which resembles the red-teaming approach to distinguish harmful user prompts [6, 26]. Furthermore, legal scholars have explored legally justifiable AI advice under UPL, attorney-client privilege, and other doctrines [32, 81, 91]. Wendel's definition of the "core lawyering functions" is a pertinent one: Important tasks like recommending the course of actions or drafting contracts cannot be delegated to AI agents due to technical limitations and accountability deficits [91]. This demonstrates how principles accumulated over centuries of legal practice now inform responsible AI systems and the call for cross-disciplinary collaborations.

Type	Permissible questions	Impermissible questions
Procedure	Can you tell me how to file a small claims action?	Can you tell me whether it would be better to file a small claims action or a civil action?
Definition	What does "certificate of service" mean?	My neighbors leave their kids at home all day without supervision. Isn't that child neglect?
Forms	I need to file for divorce and I have no idea where to begin. Is there some place I can go to find out how to get started?	The self-help divorce petition says I should list any gifts as my separate property. Should I list the money that my parents gave me last month as my separate property?
Options	What can I do if I cannot afford to pay the filing fee?	My ex-husband hasn't paid the debts that he agreed to pay in our divorce settlement. Can I be made responsible for this debt?

Table 2. Examples of impermissible questions that requires legal opinion [4]. Remarkably similar to how red-teaming in LLM development identifies harmful user inputs [6, 26], this edited list (compiled from Texas law clerk resources) distinguishes between permissible and forbidden questions Texas court personnel can answer.

Limitations and Future Research. Our study has several limitations. First, our expert sample predominantly focused on practitioners familiar with the US legal system. Ethical considerations around appropriate AI assistance may differ significantly across legal systems and cultures around the world. Second, we did not directly engage end-users like clients of legal services. Future work can specifically investigate end-user perceptions to compare and contrast with our expert-informed results. Finally, while our taxonomy conceptualizes a concrete set of dimensions, how these dimensions could change the appropriateness of AI responses remains unexplained. This may require larger-scale empirical analysis on public assessments across diverse pairings of cases and responses.

6 CONCLUSION

Today, LLM chatbots are increasingly capable of providing users with advice in a wide range of professional domains, including legal advice. However, what constitutes an appropriate AI-generated response to legal queries, where both required expertise and resulting consequences are high? To explore this, we conducted workshops with 20 legal experts using methods inspired by case-based reasoning to encourage deliberations around appropriate AI responses to legal queries in practice. Our contributions are threefold. First, we presented a set of 25 key dimensions, synthesized from expert deliberations, that impacted AI response appropriateness in the legal domain. Second, we shared experts' recommendations for AI response strategies and guiding principles for generating appropriate responses—these centered around helping users identify and prepare salient information for legal proceedings rather than to recommend specific legal actions. Finally, we posit that our case-based method has utility in engaging expert perspectives on AI response appropriateness in professional domains beyond the legal sphere. Taken together, our work sets an empirical foundation for translating domain-specific professional knowledge and practices into AI policies to steer real-world AI behavior in a more responsible direction.

7 ACKNOWLEDGEMENT

This project was supported by OpenAI's "Democratic Inputs to AI" grant program. We thank Teddy Lee, Tyna Eloundou, Alex Beutel, Aviv Ovadya, members of the grant recipient cohort, and others at OpenAI for insightful discussions and support.

REFERENCES

[1] Grievance Comm. of Bar v. Dacey, 222 A.2d 339, 347 (Conn. 1966), appeal dismissed, 386 U.S. 683, 1967.

[2] Baron v. City of Los Angeles, 2 Cal. 3d 535, 542, 1970.

[3] Cal. Bus. & Prof. Code § 6450, 2007.

[4] Legal information vs. legal advice, September 2015. URL https://www.txcourts.gov/media/1220087/legalinformationvslegaladviceguidelines.pdf.

[5] Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[7] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Jens Grossklags, Koen Hindriks, Holger Hoos, Sarit Kraus, Daan van Leeuwen, Linda van der Gaag, Qingru Liao, Chunyan Liu, Iulian Serban, Abdul Siddique, Andreas Theodorou, Francesca Toni, Albert de Waard, and Pengcheng Weng. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[8] Maria Antoniak, Aakanksha Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. Designing guiding principles for nlp for healthcare: A case study of maternal health, 2023.

[9] Roos Bakker, Romy AN van Drie, Maaike de Boer, Robert van Doesburg, and Tom van Engers. Semantic role labelling for dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 448–457, 2022.

[10] Brhmie Balaram, Tony Greenham, and Jasmine Leonard. Artificial intelligence: real public engagement. *RSA, London. Retrieved November*, 5:2018, 2018.

[11] Michael Balas, Jordan Joseph Wadden, Philip C Hébert, Eric Mathison, Marika D Warren, Victoria Seavilleklein, Daniel Wyzynski, Alison Callahan, Sean A Crawford, Parnian Arjmand, et al. Exploring the potential utility of ai large language models for medical ethics: an expert panel evaluation of gpt-4. *Journal of Medical Ethics*, 2023.

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[13] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

[14] Benjamin N Cardozo and Andrew L Kaufman. *The nature of the judicial process.* Quid Pro Books, 2010.

[15] Columbia Law School Writing Center. Organizing a legal discussion (irac, crac, etc.), 2001. URL https://www.law.columbia.edu/sites/default/files/2021-07/organizing_a_legal_discussion.pdf.

[16] Quan Ze Chen and Amy X Zhang. Case law grounding: Aligning judgments of humans and ai on socially-constructed concepts. *arXiv preprint arXiv:2310.07019*, 2023.

[17] Sasha Costanza-Chock. Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*, 2018.

[18] Alexandra D'Arcy and Emily M. Bender. Ethics in linguistics. *Annual Review of Linguistics*, 9:49–69, 2023.

[19] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23, 2023.

[20] Derek A Denckla. Nonlawyers and the unauthorized practice of law: an overview of the legal and ethical parameters. *Fordham L. Rev.*, 67:2581, 1998.

[21] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*, 2023.

[22] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[23] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, April 2021. COM(2021) 206 final 2021/0106(COD).

[24] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. Case repositories: Towards case-based reasoning for ai alignment, 2023.

[25] Robert K Fullinwider. Philosophy, casuistry, and moral development. *Theory and Research in Education*, 8(2):173–185, 2010.

[26] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[27] Sourojit Ghosh and Aylin Caliskan. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *arXiv preprint arXiv:2305.10510*, 2023.

[28] Candida M. Greco and Andrea Tagarelli. Bringing order into the realm of transformer-based language models for artificial intelligence and law. *arXiv preprint arXiv:2308.05502*, 2023.

[29] Thomas C Grey. Langdell's orthodoxy. *U. Pitt. L. Rev.*, 45:1, 1983.

[30] 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, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

[31] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[32] Claudia E. Haupt. Artificial professional advice. *Yale JL & Tech.*, 21:55, 2019.

[33] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report, 2023.

[34] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. Llm platform security: Applying a systematic evaluation framework to openai's chatgpt plugins. *arXiv preprint arXiv:2309.10254*, 2023.

[35] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. Chatgpt for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1):100105, 2023.

[36] Cong Jiang and Xiaolei Yang. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421, 2023.

[37] Albert R Jonsen. Casuistry and clinical ethics. *Theoretical Medicine*, 7:65–74, 1986.

[38] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 77–88, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593978. URL https://doi.org/10.1145/3593013.3593978.

[39] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level, 2023.

[40] Janet L Kolodner. An introduction to case-based reasoning. *Artificial intelligence review*, 6(1):3–34, 1992.

[41] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24, 2023.

[42] Davide Liga and Livio Robaldo. Fine-tuning gpt-3 for legal rule classification. *Computer Law & Security Review*, 51:105864, 2023.

[43] John Lightbourne. Algorithms & fiduciaries: existing and proposed regulatory approaches to artificially intelligent financial planners. *Duke LJ*, 67: 651, 2017.

[44] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[45] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, et al. Beyond one-model-fits-all: A survey of domain specialization for large language models. *arXiv preprint arXiv:2305.18703*, 2023.

[46] Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines, 2023.

[47] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Interpretable long-form legal question answering with retrieval-augmented large language models. *arXiv preprint arXiv:2309.17050*, 2023.

[48] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association, 2023.

[49] John Leslie Mackie. *Hume's moral theory*. Routledge, 2003.

[50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359174. URL https://doi.org/10.1145/3359174.

[51] Katherine Medianik. Artificially intelligent lawyers: updating the model rules of professional conduct in accordance with the new technological era. *Cardozo L. Rev.*, 39:1497, 2017.

[52] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: Making domain experts out of dilettantes. In *ACM SIGIR Forum*, volume 55, pages 1–27, New York, NY, USA, 2021. ACM.

[53] Jeffrey Metzler. The importance of irac and legal writing. *U. Det. Mercy L. Rev.*, 80:501, 2002.

[54] John J. Nay. Large language models as corporate lobbyists. *arXiv preprint arXiv:2301.01181*, 2023.

[55] John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. Large language models as tax attorneys: A case study in legal capabilities emergence, 2023.

[56] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, 2020.

[57] Ha-Thanh Nguyen, Wachara Fungwacharakorn, and Ken Satoh. Enhancing logical reasoning in large language models to facilitate legal applications, 2023.

[58] Gavin Northey, Vanessa Hunter, Rory Mulcahy, Kelly Choong, and Michael Mehmet. Man vs machine: how artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing*, 40(6):1182–1199, 2022.

[59] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 542–553, 2023.

[60] American Bar Association. Commission on Nonlawyer Practice. Nonlawyer activity in law-related situations: A report with recommendations. American Bar Association, 1995.

[61] Norbert Paulo. Casuistry as common law morality. *Theoretical Medicine and Bioethics*, 36(6):373–389, 2015.

[62] Denis Peskoff and Brandon Stewart. Credible without credit: Domain experts assess generative language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.37. URL https://aclanthology.org/2023.acl-short.37.

[63] Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. Effect of hierarchical domain-specific language models and attention in the classification of decisions for legal cases. In *Proceedings of the CIRCLE (Joint Conference of the Information Retrieval Communities in Europe), Samatan, Gers, France*, pages 4–7, 2022.

[64] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. Queer in ai: A case study in community-led participatory ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1882–1895, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594134. URL https://doi.org/10.1145/3593013.3594134.

[65] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

[66] Mathew Rotenberg. Stifled justice: The unauthorized practice of law and internet legal resources. *Minn. L. Rev.*, 97:709, 2012.

[67] Tulika Saha, Debasis Ganguly, Sriparna Saha, and Prasenjit Mitra. Workshop on large language models' interpretability and trustworthiness (llmit). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5290–5293, 2023.

[68] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6):887, 2023.

[69] Amy J Schmitz and John Zeleznikow. Intelligent legal tech to empower self-represented litigants. *Ohio State Legal Studies Research Paper*, (688):23, 2022.

[70] Chirag Shah and Emily M Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 221–232, 2022.

[71] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, pages 1–6, 2023.

[72] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt, 2023.

[73] Riya Sil, Abhishek Roy, Bharat Bhushan, and A. K. Mazumdar. Artificial intelligence and machine learning based legal application: the state-of-the-art and future research trends. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 57–62. IEEE, 2019.

[74] Drew Simshaw. Ethical issues in robo-lawyering: The need for guidance on developing and using artificial intelligence in the practice of law. *Hastings LJ*, 70:173, 2018.

[75] Chandan Singh, Armin Askari, Rich Caruana, and Jianfeng Gao. Augmenting interpretable models with large language models during training. *Nature Communications*, 14(1):7913, 2023.

[76] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[77] Adam Smith. The theory of moral sentiments, ed. dd raphael and al macfie, 1976. VII, iv, 34.

[78] Centaine L Snoswell, Aaron J Snoswell, Jaimon T Kelly, Liam J Caffery, and Anthony C Smith. Artificial intelligence: Augmenting telehealth with large language models. *Journal of telemedicine and telecare*, page 1357633X231169055, 2023.

[79] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative ai systems in systems and society, 2023.

[80] Tania Sourdin. Judge v robot?: Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal, The*, 41(4):1114–1133, 2018.

[81] Thomas E. Spahn. Is your artificial intelligence guilty of the unauthorized practice of law. *Rich. JL & Tech.*, 24:1, 2017.

[82] Michael Stockdale and Rebecca Mitchell. Legal advice privilege and artificial legal intelligence: Can robots give privileged legal advice? *The International Journal of Evidence & Proof*, 23(4):422–439, 2019. doi: 10.1177/1365712719862296. URL https://journals.sagepub.com/doi/abs/10.1177/1365712719862296.

[83] Sasha Fathima Suhel, Vinod Kumar Shukla, Sonali Vyas, and Ved Prakash Mishra. Conversation to automation in banking through chatbot using artificial machine intelligence language. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 611–618. IEEE, 2020.

[84] Cass R Sunstein. *Legal reasoning and political conflict*. Oxford University Press, 2018.

[85] Iddo Tavory and Stefan Timmermans. *Abductive analysis: Theorizing qualitative research*. University of Chicago Press, 2014.

[86] Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

[87] Arianna Trozze, Toby Davies, and Bennett Kleinberg. Large language models in cryptocurrency securities cases: Can chatgpt replace lawyers?, 2023.

[88] Josef Valvoda, Ryan Cotterell, and Simone Teufel. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48, 2023.

[89] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Clinical text summarization: Adapting large language models can outperform human experts, 2023.

[90] Benjamin Weiser. Chatgpt lawyers are ordered to consider seeking forgiveness, June 2023. URL https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html.

[91] W. Bradley Wendel. The promise and limitations of artificial intelligence in the practice of law. *Okla. L. Rev.*, 72:21, 2019.

[92] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84, 2021.

[93] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*, 2023.

[94] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. Disc-lawllm: Fine-tuning large language models for intelligent legal services, 2023.

[95] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[96] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.

A PROVIDED AI RESPONSE STRATEGIES AND EXAMPLES

Table 3. AI response strategies and corresponding example responses.

Response Strategies	Description	Example Response (Legal Advice)
Content violation	Displays a warning message that the user's input violates the AI's content policy.	This content may violate our content policy. If you believe this to be in error, please submit your feedback—your input will aid our research in this area.
Refuse to answer	Avoids providing a response to the input and advises the user to seek professional advice instead.	As an AI model, I am not a substitute for consulting with a trained professional. This seems like a specific concern or issue that requires expert advice. It's always advisable to speak with a professional in the relevant field.
Empathize + Refuse	Emphasizes with a user but avoids providing a response to the input.	I am sorry that this is happening to you. It sounds like a challenging situation. Unfortunately, as an AI model, I cannot provide specific legal advice. . .
Retreive (non-opinion) facts	Provides a detailed response that cites specific facts relevant to the input but avoids providing a further legal analysis.	As an AI model I cannot give legal advice specific to your situation but here is a list of laws and cases that may be helpful to look into: (1) Compliance with Applicable Laws and Regulations: Laws like the Children's Online Privacy Protection Act (COPPA) and the California Consumer Privacy Act (CCPA) may be relevant. . .
Help work through questions	Helps users to identify relevant matters in their situations and, if necessary, responds with a question requesting more information.	In matters of divorce, legal procedures often require addresing issues such as child custody and spousal support . . . . To help you with this, I will need some additional information. How many children do you have?
Recommend Actions	Analyzes a user's situation under the relevant laws and provides specific further steps that a user may take.	I regret the distressing experience of verbal abuse that you endured at school. In California, potential legal recourses may encompass claims of negligence, a violation of federal civil rights laws (Title XI) . . . As a prudent next step, I recommend initiating formal proceedings by filing a complaint with the appropriate administrative body responsible for overseeing Title IX compliance within the school or educational institution.

B WORKSHOP PARTICIPANT INFORMATION

Table 4. Workshop Participant Information

Number	Legal Experience (yrs)	Category	AI Use (General)	AI Use (Work)
P1	> 20	Law faculty	Occasional	Occasional
P2	< 5	Attorney	Occasional	Occasional
P3	> 20	Law faculty	Regular	Occasional
P4	6-10	Attorney	Regular	Regular
P5	11-15	Attorney	Occasional	Never
P6	> 20	Law faculty	Regular	Regular
P7	< 5	Law student	Regular	Never
P8	11-15	Attorney	Regular	Regular
P9	6-10	Law faculty	Occasional	Occasional
P10	< 5	Attorney	Regular	Regular
P11	< 5	Attorney	Regular	Regular
P12	< 5	Researcher	Regular	Regular
P13	6-10	Attorney	Regular	Regular
P14	< 5	Attorney	Regular	Regular
P15	< 5	Law student	Regular	Occasional
P16	6-10	Attorney	Regular	Regular
P17	16-20	Attorney	Occasional	Occasional
P18	< 5	Attorney	Occasional	Never
P19	< 5	Law student	Regular	Occasional
P20	< 5	Attorney	Regular	Regular

Note: Years of legal experience is self-reported with years of legal education removed for consistency.

C LINEAR REGRESSION OF PARTICIPANTS' AI USAGE AND DESIRED RESPONSES

Presented in Table 5, participants' receptivity to a tailored AI response is estimated by the average of the most generous answer types per each prompt. The "content warning" is marked as 0 points, the lowest comfort level, and the "recommend action" template is marked as 6 points. For example, if a participant chose both "empathize + refusal" (2 points) and "Help work through questions" (4 points) for the first case (the higher point is 4) and chose "Recommend actions" (6 points) for the second case, we marked their receptivity level as 5 points. While P13 worked on four cases, all other participants chose two cases each. The regression results (Table 6) indicate that general AI fluency significantly predicts higher comfort levels with proactive AI responses ($p < 0.05$), whereas work AI fluency is marginally associated with lower comfort levels ($p = 0.054$). The predictors explain 25.6% of variation. Further investigation is required to substantiate these preliminary relationships with a larger sample.

Table 5. AI Use and Receptivity

Number	AI Use (General)	AI Use (Work)	Receptivity
P1	1	1	1
P2	1	1	4
P3	2	1	5
P4	2	2	4
P5	1	0	4
P6	2	2	3.5
P7	2	0	5.5
P8	2	2	2.5
P9	1	1	4
P10	2	2	4.5
P11	2	2	4
P12	2	2	1
P13	2	2	3.75
P14	2	2	4.5
P15	2	1	4.5
P16	2	2	6
P17	1	1	4
P18	1	0	4
P19	2	1	6
P20	2	2	4.5

Note: A pre-survey asked participants to describe their AI usage in both professional (“Work”) and non-professional (“General”) settings, using a scale where 0 represented “Never,” 1 “Occasional use,” and 2 “Regular use.” We then estimated receptivity to more tailored responses such as opinion by averaging the most generous answer types for each case.

Table 6. Regression Results.

Predictor	Estimate	p-value
Intercept	2.4682	0.0297
AI usage in work	-0.9682	0.0543
AI usage daily	1.6773	0.0373

- Residual Std. Error: 1.199 on 17 degrees of freedom
- Multiple R-squared: 0.2557
- Adjusted R-squared: 0.1681
- F-statistic: 2.92 on 2 and 17 DF
- p-value: 0.08127