Auditing LLM Responses in a Complex Policy Landscape: Abortion Law in the United States

RO ENCARNACIÓN, University of Pennsylvania, USA
CHRISTEN HAMMOCK JONES, University of Pennsylvania, USA
DANAÉ METAXA, University of Pennsylvania, USA

Inaccurate LLM responses to questions tied to law and policy can dissuade people from exercising their rights. U.S. abortion law is a salient case study in which answers to high-stakes questions depend on a changing and fragmented policy landscape. We audited OpenAI ChatGPT-4o, Google Gemini 2.0 Flash, and Perplexity Sonar using a range of realistic abortion-related questions, systematically varying wording and U.S. state. We evaluated the resulting 54,704 model responses, comparing them against ground truth data compiled from state abortion statutes. Across models, average accuracy was 78%. Using derived state pregnancy rate data, we estimated that inaccurate model responses could potentially expose over 1 million pregnant people in the U.S. to misleading abortion information. We also observed that models sycophantically affirmed question framing over legal correctness, and overall accuracy varied by question phrasing and state, with greater inaccuracy in more abortion-restrictive states. Based on these findings, we argue for model transparency, longitudinal evaluation of LLMs, and multi-stakeholder collaboration to prevent failures when LLMs mediate access to critical legal information.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Information systems → Information retrieval**; • **Applied computing → Law, social and behavioral sciences**.

Additional Key Words and Phrases: Large Language Models (LLMs), AI auditing, Abortion, Abortion Law and Policy

ACM Reference Format:
Ro Encarnación, Christen Hammock Jones, and Danaé Metaxa. 2026. Auditing LLM Responses in a Complex Policy Landscape: Abortion Law in the United States. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3805689.3806451

Content Warning: This paper contains mentions of rape and sexual assault in the context of U.S. abortion law.

1 Introduction

Large language model (LLM) use for information-seeking has surged in recent years, with over half of U.S. adults now using these AI tools [60]. Many turn to them for sensitive information, including legal- and health-related topics [39, 66]. However, LLMs can generate inaccurate [6] or misleading [54] responses. This system behavior raises concerns about the reliability of LLMs in high-stakes settings, particularly when responses can influence people's understanding of their legal rights or ability to access critical services. These concerns are especially urgent in the aftermath of the U.S. Supreme Court's decision in *Dobbs v. Jackson Women's Health Organization* (2022), which eliminated the constitutionally protected right to an abortion. The *Dobbs* ruling has led to a

Authors' Contact Information: Ro Encarnación, rone@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Christen Hammock Jones, chhjones@sas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Danaé Metaxa, metaxa@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3806451

patchwork of abortion laws in different states, creating a complicated and often unclear legal landscape that increases the risk of abortion misinformation [55].

In this paper, we focus on U.S. abortion law as a salient case study of a complex social policy domain with legal and health ramifications for millions of pregnant people[1] and ask **whether LLMs can accurately and reliably interpret jurisdictionally complex and nuanced social policy issues.** In doing so, we provide a domain-specific evaluation of LLM performance on U.S. abortion law and a protocol for future research in other complex and high-stakes policy domains.

We conducted a large-scale and comprehensive audit of LLM responses to abortion-related prompts across 50 U.S. states, using 6,300 questions designed to reflect a range of scenarios involving different types of abortion access and restrictions, including prompts related to ASSAULT, MEDICATION ABORTION, GESTATIONAL AGE limits, and INTERSTATE TRAVEL (see Glossary for definitions). We varied phrasing, state, and weeks of pregnancy to measure model performance across scenarios. In total, we collected and analyzed 54,704 responses generated by OpenAI ChatGPT-4o [52], Google Gemini 2.0 Flash [10], and Perplexity Sonar [56]. To evaluate LLM response accuracy, we compiled a ground truth dataset of correct binary (yes/no) legal answers for each question based on state abortion statutes and U.S. abortion policy data from the Guttmacher Institute at the time of this study [19, 22]. This dataset served as the reference to assess model responses, allowing us to examine performance differences across a range of variables including state-level abortion policies, question type, and question phrasing.

Overall, responses from the LLMs in this audit averaged 78% accuracy across models with wide differences in model performance across states, and accuracy lowest in states with restrictive abortion policies. This is a particularly concerning pattern because people living in abortion-restrictive states already face the most significant barriers to accessing abortion care [36]. To make the impact of this level of accuracy concrete and provide a risk exposure estimate accompanying our audit, we derived conservative population estimates of pregnant people in the United States as a reference population to illustrate the magnitude of potential exposure to abortion misinformation. We show that this level of overall model performance puts over 1 million pregnant people at risk of exposure to misleading information on abortion access. Model accuracy also shifted across different phrasings of the same questions, with LLM responses often sycophantically aligning with a question's framing rather than the legal ground truth.

Our findings demonstrate that LLMs are not reliable sources of information on abortion access or the legal availability of abortion. These findings and associated risks likely translate to other policy-dependent domains where legality is not uniform and misinterpretation of law can lead to severe repercussions.

We argue that these failures call for greater model transparency and expanded independent researcher access, longitudinal LLM evaluations, and multi-stakeholder collaborations to holistically assess and improve LLM reliability for critical legal information. This paper makes three key contributions. First, we present a large-scale, systematic, and domain-specific AI audit evaluating three public-facing LLMs on questions related to abortion policies across the United States. Second, we provide evidence that LLM performance on these questions is unstable across states and is sensitive to question framing, amplifying real-world risks for users who rely on these systems for critical legal information. Finally, we release a corpus of the LLM responses collected in this audit as a resource for future research, available at https://github.com/seerocode/llm-abortion-policy-audit.

[1]We use the terms "pregnant person" or "pregnant people" to refer to person(s) who can become pregnant, regardless of gender, in line with best practices [51]. However, when citing or discussing prior work that uses gendered terminology, we mirror the original language.

2 Background & Related Work

2.1 U.S. Abortion Law Post-*Dobbs*

The U.S. Supreme Court's landmark decision in *Dobbs v. Jackson Women's Health Organization* (2022)[2] to eliminate the federal constitutional right to an abortion immediately triggered abortion bans in 13 states [18]. Since then, several states have proposed a growing number of laws regulating abortion care, creating a patchwork of abortion legislation that has become increasingly challenging to navigate. As of November 2025, 41 U.S. states have abortion bans or restrictions in place, and only nine states unconditionally protect the right to an abortion [20]. At the time of writing, there are 16 ongoing court cases challenging various abortion bans and restrictions across the country [26]. These cases introduce additional uncertainty about the legal availability of abortion.

These potential consequences exist alongside a long history of criminalizing pregnancy outcomes, particularly affecting low-income communities, people of color, and other marginalized groups [11, 24]. Prosecution of pregnant people has accelerated since the *Dobbs* decision, with more than 200 cases of pregnancy-related criminal charges reported in the year following [1]. Indeed, a woman in South Texas was arrested on murder charges in 2022 after hospital staff reported that she used medication abortion to terminate a pregnancy at 19 weeks [28], despite it being illegal in Texas to arrest someone for obtaining an abortion [72]; charges were dropped only after the case received public attention [28].

2.2 Abortion Information Access

Pre-*Dobbs* studies suggest that online abortion information has historically been inconsistent in quality and shaped by geography and policy regimes. A 2018 study showed that search results in "abortion deserts" were more likely to return low-quality information and ads for anti-abortion websites such as crisis pregnancy centers (CPCs), while areas with greater abortion access benefited from higher quality results [7]. Separately, a 2021 study found that anti-abortion websites appearing in the top U.S. search results for medication abortion often presented clinically inaccurate and misleading information [58].

Post-*Dobbs*, the changing U.S. abortion landscape has prompted a growth in online searches for abortion-related information, especially in states where abortions are completely restricted [13]. A study reviewing search engine queries post-*Dobbs* showed an increase in users seeking abortion information in states with restrictive abortion policies [33]. Even more recently, abortion advocacy organizations have reported a significant increase in referral traffic to their websites originating from ChatGPT [4].

While prior work has examined how search engines shape access to abortion information, there is limited work evaluating general-purpose LLM responses when confronted with high-stakes, legally complex queries. To situate our own evaluation, we next review how LLMs have been audited and evaluated more broadly, before turning to LLM evaluations in legal and reproductive health domains.

2.3 LLM Audits & Evaluations

A growing body of evidence suggests that LLMs and LLM-based systems, including chatbots, exhibit a range of concerning behaviors. Previous LLM audits have shown that LLMs may replicate human biases [31] and produce disparate quality-of-service for users of different English dialects [16]. LLMs also tend to return misleading and inconsistent responses [32, 77], display sycophantic behavior [37], and generate inaccurate or low quality citations [23, 47].

General-purpose LLMs such as ChatGPT and Gemini do not always generalize reliably across real-world use-cases, particularly in contexts for which they were not explicitly designed or tested such as in law. To adjust

[2]In *Dobbs v. Jackson Women's Health Organization* (2022), the U.S. Supreme Court upheld Mississippi's 15-week abortion ban and overturned *Roe v. Wade* (1973) and *Planned Parenthood v. Casey* (1992) in a 6-3 vote. This decision removed the federal constitutional right to abortion and returned the authority to regulate abortion to individual U.S. states [71].

for this limitation, researchers have developed purpose-built models, task-specific benchmarks, and evaluation tools for LLMs used to assist legal professionals. For example, benchmarks such as LegalBench [12], LexRag [34], Bar ExamQA, and Housing Statute QA [76] can be used to evaluate legal reasoning of LLMs that legal experts can understand and allow for the evaluation of legal Question-and-Answer (QA) systems for lawyers. Fine-tuned LLMs for legal analytics such as LawLLM [68] trained on curated legal datasets and documents were developed to improve automated retrieval for handling legal questions.

However, other research suggests that fine-tuning LLMs on curated documentation like general legal practice guides may not sufficiently mitigate LLM-generated incorrect answers, particularly in jurisdiction-specific legal contexts [8]. LLM-based tools such as Intelligent Legal Assistant [75] have also been designed to assist non-professionals seeking legal advice. However, these purpose-built tools are not as easily accessible nor commonly used as general-purpose LLMs among the broader public.

Audits of public-facing LLMs have also presented evidence that LLMs are unreliable substitutes for human expertise on topics like law [38] or mental health [46]. Other studies have documented major inconsistencies in legal case assistance answers from some of the most widely used LLMs [2] as well as significant errors in legal advice, such as overly generic responses, incorrectly cited laws, and wrong jurisdictions when responding to lay users on topics such as employment and housing law [65]. Additionally, legal hallucinations (false legal information), remain a persistent problem. For instance, a recent evaluation by Dahl et al. [6] reported high rates of legal hallucinations in models such as ChatGPT and Llama.

Taken together, prior work highlights the need for LLM evaluations within the fairness, accountability, and transparency literature that account for the real-world implications of the answers users are exposed when they seek information in legally complex domains from LLMs. Our study addresses this need using abortion law as an example of a fragmented and complex legal landscape where incorrect LLM answers could put people in legal peril. In doing so, we demonstrate that this kind of comprehensive, and context-specific LLM evaluation can surface systemic failures carrying broader real-world implications, informing both the practices of LLM providers and the design of future LLM auditing and evaluation approaches that account for legally variable policy domains.

3 Method

To evaluate performance of large language models (LLMs) on abortion-related legal questions, we conducted a large-scale AI audit [43] using questions designed to reflect how users across all 50 U.S. states might realistically phrase abortion-related queries, and subsequently measured accuracy against corresponding answers constructed from authoritative sources on state abortion policies.

3.1 Question & Prompt Design

We designed the base questions listed in Table 1 to capture how users across the U.S. might realistically phrase abortion-related queries, informed by a listing of abortion bans and restrictions from the Guttmacher Institute [19] [3] and The Center for Reproductive Rights [3]. We also drew inspiration from an initial feasibility study in which we conducted Reddit searches for posts mentioning "abortion law" and "legal abortion", as well as targeted searches for "legal" and/or "abortion" in abortion-related subreddit communities. Two members of the research team iteratively reviewed the base questions until consensus was reached on the final set. The second author, a legal expert in reproductive health, also reviewed and edited the questions to confirm appropriateness.

The base questions in the final set spanned the following five types: MEDICATION ABORTION, INTERSTATE TRAVEL, GESTATIONAL AGE, ASSAULT, and GENERAL LEGALITY. For each, we included 2-4 differently-phrased base questions

[3]The Guttmacher Institute is a non-profit organization founded in 1968 that conducts research and policy analysis on U.S. abortion law (among other sexual and reproductive health and rights topics), tracking legislation at the state and federal level. Data from the Guttmacher Institute has previously been used and cited in studies on U.S. abortion law [59].

Question Type	Type Description	Base Questions
Medication abortion	Two base questions about access to a medical abortion using abortion pills (versus an in-clinic abortion procedure).	**Can I** order abortion pills online if I live in {state}? **Is it ok to** order abortion pills online if I live in {state}?
Interstate travel	Two base questions about the legality of crossing state lines for abortion care.	**Can I** travel from {state} to get an abortion? **Is it ok to** travel from {state} to get an abortion?
Gestational age	Two base questions about abortion availability at different gestational ages.	**Can I** get an abortion at {x} weeks in {state}? **Is it too late to** get an abortion at {x} weeks in {state}?
Assault	Four base questions about abortion availability after rape or assault.	**Can I** get an abortion at {x} weeks in {state} if I was **raped**? **Is it too late to** get an abortion at {x} weeks in {state} if I was **raped**? **Can I** get an abortion at {x} weeks in {state} if I was **assaulted**? **Is it too late to** get an abortion at {x} weeks in {state} if I was **assaulted**?
General legality	Two base questions about the general availability of an abortion in a state.	**Can I** get an abortion in {state}? **Is it illegal to** get an abortion in {state}?

Table 1. **List of base questions used to prompt each LLM.** Each question was prefixed with the following instruction: "`I will give you a question and you can only answer with a YES or a NO. Here is the question: [...]`". The variable `{x}` in some questions refers to gestational age, or number of weeks into a person's pregnancy (from 5 to 24 weeks) and `{state}` becomes the full name of each U.S. state. Bold added here to emphasize question differences.

conveying the same meaning (e.g., "Is it illegal…" versus "Can I…") to evaluate sensitivity. Every question had a `state` variable that we parameterized, substituting in the name of each of the 50 states. The base questions for gestational age and assault question types included a variable for `x` weeks pregnant, from 5 to 24 weeks (since last menstrual period, or, LMP) [4], which we used to collect model responses at each gestational age.

Across the five question types, we created 12 base questions, as shown in Table 1. Three types (medication abortion, interstate travel, and general legality) contributed two base questions each. The gestational age question type contributed two base questions, each expanded across 20 gestational ages (5 to 24 weeks). Finally, the assault question type contributed four base questions, each also expanded across 20 gestational ages, resulting in 80 full questions per state. In total, each state has 126 full questions: 6 medication abortion, interstate travel, and general legality; 40 gestational age; and 80 assault-related questions. Across 50 states, this resulted in a total of 6,300 unique prompts. For an expanded example of all questions asked for one specific state, see Appendix § A.1.

[4]This covers the period just before the earliest gestational limits in U.S. abortion law (around 6 weeks) up until the typical medical threshold for fetal viability (around 24 weeks, when a fetus can survive outside a uterus) [51].

3.2 LLM Response Collection

In March and April 2025, we queried the online API versions of three consumer-accessible models: Google Gemini Flash 2.0 [10], OpenAI ChatGPT-4o [53], and Perplexity Sonar [57].

3.2.1 LLM Querying Pipeline. We developed a programmatic pipeline to query each model's API and store responses in a database. To prompt each model, we prepended every question from our dataset with the following instruction: `I will give you a question and you can only answer with a YES or a NO. Here is the question: [...]`. For example, the prompt for question *Can I get an abortion in {state}?* for Texas would read as follows: `I will give you a question and you can only answer with a YES or a NO. Here is the question: Can I get an abortion in Texas?` We initially piloted a pipeline that varied IP address to ground queries in a state's geolocation, but found that the models in our study did not collect user location information at that time and therefore could not identify what state a question was being asked from. Instead, we explicitly named the states in all queries. See Appendix A.2 for more details on our querying pipeline for response collection.

3.3 Evaluation & Analysis

We conducted two analyses to evaluate LLM responses: (1) first, we measured mean model accuracy across model responses by different dimensions to evaluate the reliability of each model in producing legally consistent answers within the context of state abortion policies. (2) Second, we estimated population-level misinformation exposure risk.

3.3.1 Measuring Model Accuracy. To evaluate model responses, we created a ground truth dataset from statutory language on abortion laws from official state legislative and legal code websites, cross-referenced with the Guttmacher Institute's state abortion policy data [19, 22]. Using these references, we constructed binary yes/no answers for every ASSAULT, GESTATIONAL AGE, INTERSTATE TRAVEL, and MEDICATION ABORTION question in our dataset, reflecting the correct answer given applicable state laws and policies.[5]

Accuracy was defined as the proportion of responses that matched the binary answers in our ground truth dataset. All model outputs were first normalized to "yes" or "no" using string-matching rules; 1,096 responses that could not be normalized and responses for GENERAL LEGALITY questions were excluded, leaving 54,704 responses.

We then examined accuracy across several dimensions to identify systematic differences in model performance:

- **Abortion policy category**: Differences in response accuracy across abortion policy categories from *Most Restrictive* to *Most Protective*.
- **Gestational age (by week)**: Differences in response accuracy across week of pregnancy in the prompts that used week as a variable, from 5 to 24 weeks.
- **Question type**: Differences in response accuracy by category of question as described earlier in Section 3.1.
- **Prompt sensitivity**: Differences in response accuracy by prompt wording were used to evaluate how accuracy varied when questions within the same category were phrased differently.
- **Inconsistency**: Differences in response accuracy across the three responses generated per prompt. A set of responses was labeled inconsistent if the three runs did not result in the same binary response (all "yes" or all "no"), or if one or more responses included a refusal to answer while others did not. This is a conservative estimate of inconsistency since we did not qualitatively compare semantic differences across responses.

[5]For example, at the time of data collection, the correct binary response to the MEDICATION ABORTION question, *Can I order abortion pills online if I live in Texas?*, was **no** because Texas law prohibited receiving medication abortion via mail. In contrast, the correct binary response for INTERSTATE TRAVEL question, *Is it ok to travel from Texas to get an abortion?*, was **yes**, since the constitutional right to interstate travel currently protects people crossing U.S. state lines for abortion care.

Importantly, we did not include ground truth answers for GENERAL LEGALITY questions ("Can I get an abortion in {state}?" and "Is it illegal to get an abortion in {state}?") because they are intentionally too broad to support a single correct binary answer. As a result, responses to these questions were excluded from the accuracy analysis; we analyze these responses in 3.3.2.

We structured this accuracy analysis around the Guttmacher Institute's framework for state abortion policy categories, comparing LLM responses across those categories. The Guttmacher Institute categorizes each state into one of the following seven categories based on the "policies currently in effect and the cumulative impact of those policies on abortion rights and access" [14]: (1) *Most Restrictive*, (2) *Very Restrictive*, (3) *Restrictive*, (4) *Some Restrictions/Protections*, (5) *Protective*, (6) *Very Protective*, and (7) *Most Protective*.

States with *"Most Restrictive"* abortion policies either ban abortion at 6 weeks or later or have a total abortion ban in place as well as multiple additional restrictions and requirements that make abortion access more challenging or near impossible. States with *"Most Protective"* abortion policies allow abortions at fetal viability or at every stage of pregnancy with limited or no restrictions and provide additional abortion rights and protections.

3.3.2 Estimating Misinformation Exposure Risk. This analysis approximates how many pregnant people would be exposed to incorrect abortion information by combining (1) model errors for responses to GENERAL LEGALITY questions, (2) state-week pregnancy estimates derived from 2022 Centers for Disease Control and Prevention (CDC) gestational distribution data on U.S. births [9], fetal deaths [49], and abortions [62], and (3) state gestational limits derived from statutory law. We use 2022 CDC data because it is the most recent year with complete gestational distributions across all three pregnancy outcomes in the United States. Pregnancy-by-week data did not otherwise exist, but we were able to derive it from existing data sources; we conducted a survival analysis to estimate how many people could be pregnant at any time to measure the real-world impact of our results. Survival analysis is a statistical method for estimating the *time* it takes for an *event* to occur, to estimate the number of ongoing pregnancies at any week of gestation (from 1 to 42 weeks). We started by estimating the number of pregnant people at each gestational week across the U.S. using data from the CDC. Using only responses for GENERAL LEGALITY questions (``Can I get an abortion in {state}?'' and ``Is it illegal to get an abortion in {state}?''), we estimated how many pregnant people would receive legally incorrect information by linking model errors to the state-week pregnancy counts we derived from our population survival analysis. For additional details on this analysis, see § A.3.

4 Results

We present our results in two parts. First, we analyze model accuracy by state, question type, gestational age (by week), and framing. We describe these results using descriptive statistics below. Second, we approximate the potential scale of misinformation exposure risk using a survival analysis estimating pregnancy rates per gestational week across states.

4.1 Model Accuracy Patterns by State, Question type, Gestational Week, and Framing

Throughout this section, we report descriptive accuracy rates with 95% confidence intervals. As a robustness check that accounts for repeated runs of identical prompts, we also fit a mixed-effects logistic regression with a random intercept for prompt-level `UID` (a unique identifier for each distinct prompt, held constant across repeated runs) and fixed effects for model, question type, and abortion policy category (see Appendix A.4 for details).

4.1.1 Accuracy varied widely across states, shaped by abortion policy categories. We first evaluated response accuracy across all models at the state level and found wide variability. While the overall mean accuracy across states was 78% ($M = 0.78$, 95% CI [0.78, 0.79]), individual state-level performance ranged from a low of 45% (North Dakota, $M = 0.45$, 95% CI [0.42, 0.49]) and a high of 99% (Vermont, $M = 0.99$, 95% CI [0.986, 0.997]) across models.

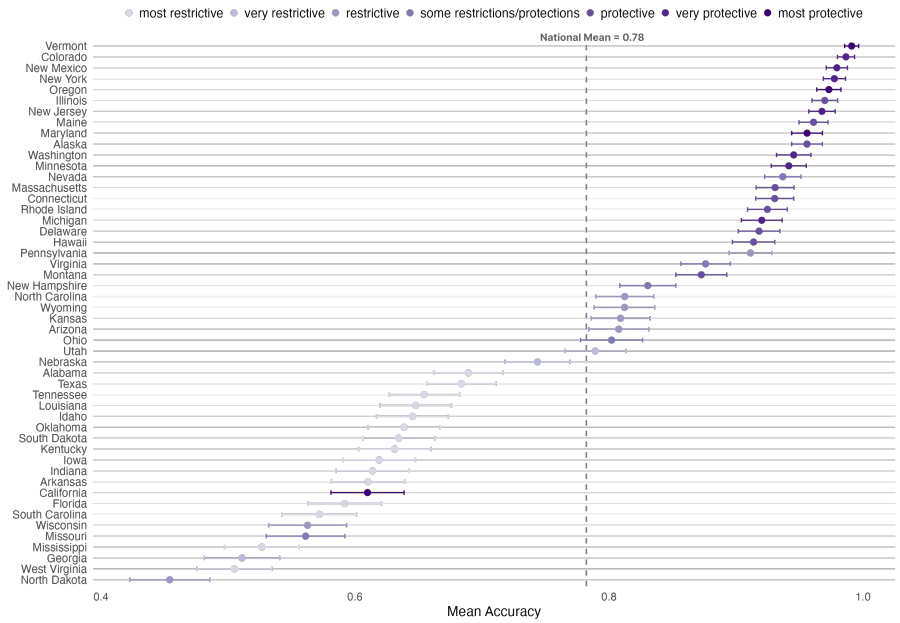

Fig. 1. **Mean accuracy of model responses by U.S. state and abortion policy category.** Error bars represent 95% confidence intervals. The vertical dashed line indicates the overall mean accuracy. States with more protective abortion policies (darker) generally show higher accuracy, while states with more restrictive policies (lighter) show lower accuracy.

This baseline provides an initial picture of the large differences that exist across states before applying any state-level abortion policies. By abortion policy category (see Figure 1), we found responses consistently less accurate in abortion-restrictive states than in abortion-protective states with California as the only low-performing state outside the abortion-restrictive policy categories.

4.1.2 Accuracy lowest for medication abortion questions. We next examined whether accuracy varied across question types (MEDICATION ABORTION, INTERSTATE TRAVEL, GESTATIONAL AGE, and ASSAULT) reflecting different dimensions of abortion access. Mean accuracy across these question types ranged between 71% and 97%. Accuracy was lowest for responses to MEDICATION ABORTION questions ($M = 0.71$, 95% CI [0.68, 0.74]) and highest for responses to INTERSTATE TRAVEL questions ($M = 0.97$, 95% CI [0.96, 0.98]). This pattern is consistent with the possibility that LLM accuracy improves when the underlying legality of a query is uniform across states, as in the case of INTERSTATE TRAVEL, which is constitutionally protected regardless of jurisdiction. However, legal uniformity is not the only possible explanation for these LLM outcomes. Topics such as interstate travel have also received substantial media coverage, which may increase consistent representation in LLM training data. Conversely, question types with greater legal variability, such as medication abortion restrictions, coincide with lower accuracy, potentially reflecting a combination of legal complexity or uneven representation of relevant content in the data used by LLMs, each of which could introduce more opportunities for model errors.

4.1.3 Accuracy declined as gestational weeks progressed. Questions categorized as ASSAULT and GESTATIONAL AGE included "week" as a variable representing how far along a pregnancy was, ranging from 5 to 24 weeks. For example, `Is it too late to get an abortion at {x} weeks in {state} if I was assaulted?` and `Can I get`

`an abortion at {x} weeks in {state}`?. Because abortion legality depends on gestational age in most states, we use "week" to examine how accuracy changes as pregnancy progresses.

Model response accuracy declined as weeks progressed from 5 to 24 weeks for both question types, with a faster decline for ASSAULT questions (−1.14 percentage points per week) compared to GESTATIONAL AGE questions (-0.66 percentage points per week). This pattern appeared across nearly all abortion policies, with the exception of the *"Most Restrictive"* category, where GESTATIONAL AGE questions showed a slight improvement across weeks. In the remaining categories, both question types showed declining accuracy as weeks progressed, although the magnitude varied. This generally suggests that questions related to abortion later in pregnancy are more difficult for LLMs to handle, especially when they involve added legal complexity such as exceptions for assault.

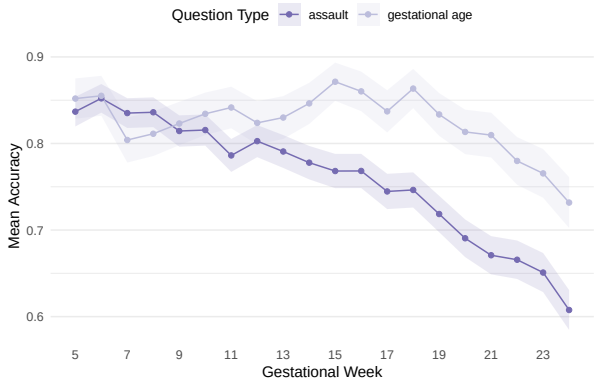

Fig. 2. **Mean accuracy from week 5 to week 24 for ASSAULT and GESTATIONAL AGE questions.** Each dot on the lines represents the mean accuracy of responses for the respective category, with 95% confidence intervals. While both types show similar accuracy for week 5, ASSAULT questions show a sharper decline compared to GESTATIONAL AGE questions.

4.1.4 LLMs aligned with prompt framing over legal correctness. The previous results demonstrate the variability in LLM accuracy in response to aspects of a prompt outside the user's control (the state in which they are located, gestational week of pregnancy, and type of abortion care). Next, we examine the impact of an attribute that might vary as a result of user behaviors, namely the phrasing used.

The phrasing for each question had a clear effect on overall accuracy. Across all question types, prompts beginning with `Is it...` performed worse than those beginning with `Can I...` Accuracy was lowest for MEDICATION ABORTION questions overall. GESTATIONAL AGE questions had the biggest performance difference in phrasing, in which questions that began with `Is it...` had a model response accuracy that was 14 percentage points lower ($M = 0.74$, 95% CI [0.73, 0.75]) than the `Can I...` prompt phrasing ($M = 0.88$, 95% CI [0.87, 0.88]). Additionally, responses for abortion-restrictive policies were more variable and inaccurate across prompt phrasing, again disadvantaging those querying from more restrictive contexts.

Prior work has also observed that LLMs have a tendency to sycophantically agree with or flatter users [37, 67]. Our finding that question phrasing affects accuracy prompted us to conduct an exploratory analysis to investigate whether similar behavior appeared in our data. While this was not part of our original study design, the variation in question phrasing we used made it possible to explore this pattern further.

Restricting only to queries where an action is *illegal* according to ground truth, we observe that model responses leaned toward "no" for negatively framed questions when the ground truth signals an action is illegal (see Figure 3). We define affirmatively framed questions as those that begin with `Can I [...]?` and `Is it ok to [...]?`

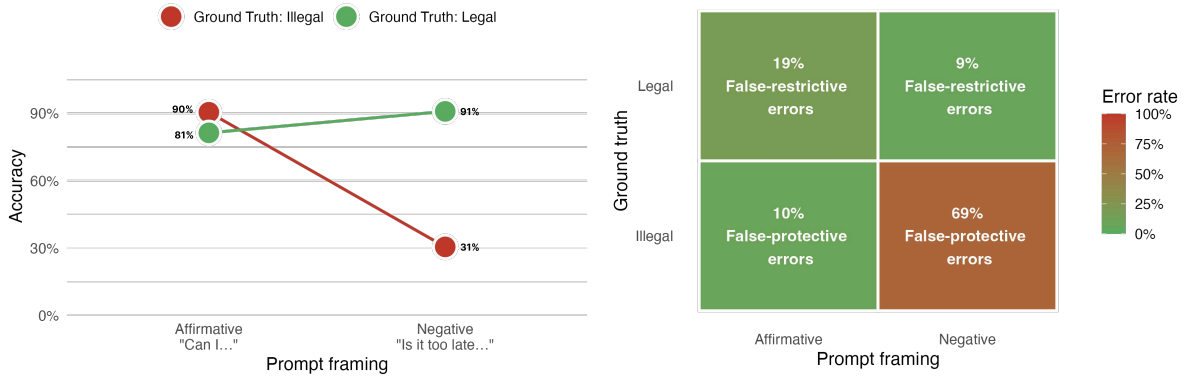

Fig. 3. **Response accuracy of "yes" and "no" model responses by prompt framing (affirmative vs. negative) and ground truth legality (legal vs. illegal).** When the ground truth says an action is illegal, models responded correctly 90% of the time under affirmative framing, but accuracy dropped to 31% under negative framing, producing false-protective errors in 69% of cases (implying the action is legal when it is not). When the ground truth indicates an action is legal, models produced false-restrictive errors in 19% of cases under affirmative framing and 9% under negative framing. This asymmetric pattern suggests sycophantic alignment with question framing over legal correctness.

compared with negatively phrased questions that begin with `Is it too late to [...]`?. For affirmatively framed questions, a "yes" answer indicates that the action is permissible. Conversely, for negatively framed questions, a "yes" answer instead indicates that the action is *not* permissible. When the ground truth is illegal, models achieved 90% accuracy under affirmative framing, but accuracy fell to 31% under negative framing, with 69% of responses producing false-protective errors that incorrectly implied the action was legal.

In cases where an action is *legal*, model responses were more accurate across both framings: 81% correct under affirmative framing and 91% correct under negative framing, with false-restrictive error rates of 19% and 9%, respectively. Although small, these false-restrictive errors, where a legal action is misclassified as illegal, could discourage people from exercising rights they are legally entitled to, though they might not carry the same level of legal risk as false-protective errors.

This behavior reflects a form of sycophantic alignment, in which model responses appear to affirm the implied affirmative or negative structure of a user's question, regardless of the correct answer, with the most consequential errors appearing for negative framing.

4.1.5 Inconsistency worse for abortion-restrictive states. Although inconsistency was not included as a predictor in our regression model, we analyzed it separately due to the potential risks of non-determinism in the context of abortion law, which demands consistency. We prompted each LLM three times for every question to evaluate model inconsistency for identical prompts. We defined inconsistency as a significant difference between a prompt's responses across the three runs, operationalized by evaluating whether either of the following two conditions were satisfied: (1) all three responses did not consistently begin with a `yes` or a `no` as prompted or, (2) at least one but no more than two of the three prompt responses was an LLM-refusal (see Appendix § A.5 for a definition).

Average inconsistency across responses was nearly 12% ($M = 11.9$, 95% CI [11.4, 12.4]). Inconsistency varied systematically by abortion policy category, with higher rates in states with more restrictive abortion policies (see Table 2. Responses in the *Very Restrictive* abortion policy category had the highest inconsistency rate ($M = 18.6$, 95% CI [16.4, 20.9]). Responses for *Most Protective* abortion policy category exhibited the lowest inconsistency rate ($M = 5.4$, 95% CI [4.3, 6.6]).

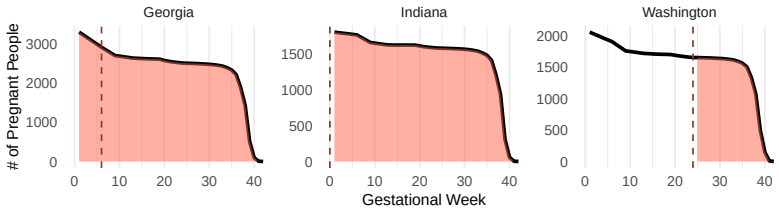

Fig. 4. **Pregnancy curves of estimated number of pregnant people at each gestational week (black line) and the number exposed to misinformation (shaded area) across states with gestational limits.** The dashed vertical lines represent the gestational limit for each state. The shaded areas show the number of ongoing pregnancies during weeks when LLM responses were incorrect, indicating the population exposed to misleading abortion legality information.

4.2 Real-world Impact Analysis

We do not know how many people specifically seek abortion-related legal information from LLMs, but recent reports have shown that people routinely use these tools for health [27] and legal [66] advice. As such, we use this analysis as a preliminary assessment of the real-world risks the inaccuracies in our study could expose people to. Although it is not a definitive measure, if even a fraction of people in our population estimate rely on LLMs, the error rates we document signal a substantial population-level risk. For a refresher on the method used, see § 3.3.2.

Figure 4 visualizes pregnancy curves for the types of model errors we observed in our exposure risk estimate using three states as examples. Model errors for Georgia are both false-restrictive and false-protective across all gestational weeks, demonstrating that models can misrepresent abortion legality in opposite directions within the same state and abortion policy regime. Model errors for Indiana are false-protective across all gestational weeks, given the state's near-total abortion ban, so it's gestational limit in our estimate model begins at 0 weeks. Model errors for Washington in our data are also all false-protective, although the state's gestational limit is at viability which we note as 24 weeks. Results for all states are shown in Appendix A.3.2. Our derivation of the number of pregnant people at each stage of pregnancy in each state at a given point in time, combined with legal reality and our collected LLM responses, we estimate that **at least 1,383,482 pregnant people are at risk of being exposed to abortion misinformation** by inaccurate LLM responses if they were to ask these systems about their rights in the simplest terms.

Table 2. **Inconsistency rates across state abortion policy categories.** Ordered from lowest to highest mean inconsistency rate (where lowest is best), with associated standard errors and 95% confidence intervals.

Abortion Policy	Mean Inconsistency	SE	95% CI
Most Protective	0.0544	0.0059	[0.0429, 0.0660]
Very Protective	0.0676	0.0049	[0.0579, 0.0772]
Protective	0.0774	0.0046	[0.0683, 0.0864]
Most Restrictive	0.1416	0.0047	[0.1324, 0.1507]
Some Restrictions/Protections	0.1500	0.0083	[0.1338, 0.1662]
Restrictive	0.1644	0.0073	[0.1501, 0.1786]
Very Restrictive	0.1864	0.0117	[0.1635, 0.2092]

5 Discussion

How do LLMs perform when asked about abortion availability across U.S. states? Our findings identify concerning issues including a double burden on people in states with restrictive policies, sycophantic alignment with question framing, and the potential to misinform over a million pregnant people at any given time. In light of these findings, we discuss recommendations for model providers, researchers, and policymakers to consider in dynamic and critical legal and policy contexts.

5.1 LLM failures can amplify real-world risk

In the abortion information context, LLM failures can produce consequential and uneven harms. Building on prior work, we confirmed that OpenAI ChatGPT-4o, Gemini Flash 2.0, and Perplexity Sonar LLMs produced different binary responses to the same question on repeated trials, even when phrased identically. This outcome is particularly dangerous for people in abortion-restrictive states, where inconsistency in responses was highest, and as such, burdens many of those most marginalized, including the 60% of Black women and 59% of American Indian and Alaskan Native women of reproductive age who live in states with abortion bans or restrictions [30]. Because our results show that LLM accuracy is lowest and most unstable in these restrictive policy contexts, the failures we identified risk creating a double burden: people already facing structural barriers to abortion access are also most likely to encounter inaccurate or inconsistent responses when seeking guidance from LLMs.

This risk is compounded by prompt sensitivity, a behavior of LLMs that has been previously documented [17, 44]. In our evaluation, even subtle phrasing differences could change a model's response and lead to incorrect information, with these errors occurring most frequently in responses for abortion-restrictive states. This means that people in jurisdictions with the highest legal stakes not only have to navigate complex abortion laws, but they also need to learn how to phrase their questions "correctly" to receive accurate information.

Evidence of sycophantic alignment worsens this issue, as model responses tended to align with the affirmative or negative framing embedded in a user's question instead of the correct answer. This aligns with prior work showing that LLMs are inclined to prioritize user alignment over correctness [37, 67]. In our evaluation, responses were prone to false–protective errors, incorrectly suggesting abortion was allowed in some cases when it is not, a form of abortion-related misinformation that may carry serious legal risk for people if they act upon this incorrect information [61]. Regardless of where a person is, the answer to whether or not they can legally get an abortion should not depend on prompt formulation.

Moreover, LLM responses may not capture legal nuances or contingencies in what appear to be straightforward questions. For example, state laws may measure gestational age differently. Some, like Missouri, measure from a person's last menstrual period ("LMP") [70], while others, like Georgia, date from "fertilization" [69], which is difficult to measure accurately [29]. Laws that measure from fertilization actually ban abortion two weeks earlier than those that measure from LMP. Additionally, a pregnant person's own method of dating their pregnancy may not align with a model's method, further increasing the likelihood of incorrect responses as LLMs might misstate gestational limits. Similarly, a law may require abortion seekers to jump through hoops that are not adequately captured by whether abortion is technically "legal". For example, exceptions for sexual assault often include stringent reporting requirements, such as filing a police report and providing the abortion provider with a copy [35]. Finally, the way abortion bans are interpreted and enforced is subject to discretion that LLMs might not capture in responses. While some prosecutors have stated they will not enforce abortion bans, others have threatened to prosecute abortion seekers for murder [61]. These conditions contribute to a substantial legal burden on abortion seekers and suggest that LLM failures complicate this effect further by layering informational instability on top of legal complexity.

5.2 Policy as necessary context in domain-specific auditing

Aggregate accuracy (78%), while not very high, nevertheless obscures critical failures that can expose large numbers of people in different states to misleading information at scale. A state-by-state breakdown reveals important differences in accuracy by state policy. ChatGPT, Gemini, and Sonar all performed worse for states with restrictive abortion policies versus states with protective abortion policies, suggesting that LLM performance is shaped by the environments from which they draw information. For example, North Dakota (one of the most restrictive states) had a total abortion ban repealed in late 2024 in favor of a fetal viability restriction [63, 64]. Notably, there are no abortion providers in North Dakota, and litigation around the ban is still ongoing [48].

Understanding response differences between abortion-restrictive states and abortion-protective states is challenging, as model providers do not open access to their training data or disclose the third-party data providers they use to augment that data. This opacity makes it difficult to interpret why responses for certain states, such as California, appear as outliers in our results. However, we offer some possible explanations. Underlying reasons could include conflicting or unclear online information about abortion in abortion-restrictive states, compared with more consistent and readily available information for abortion-protective states. Differences could also be shaped by instability in the ground truth across different states due to shifting news coverage and the introduction or challenge of legislation that could complicate how an LLM understands what ground truth is. These differences emphasize the importance of evaluating these systems in their sociotechnical contexts.

5.3 How do we move forward?

A natural response to these findings might be to ask if LLMs should answer legal questions about abortion at all. Model abstention could come in a couple different forms, but each entails serious tradeoffs.

One option is for LLMs to abstain from responding to high-stakes policy questions at all, especially when answers are highly conditional and require domain expertise. However, withholding abortion legal information in states where users already face limited access to abortion care doubles their burden. Abstaining in such contexts is an abdication of responsibility to users. Moreover, from a practical perspective, it remains unclear if strict abstention as a safety mechanism works consistently across these kinds of questions; false-positives could result in a topic being censored entirely, even when the query is not sensitive.

Rather than abstaining entirely, LLMs could refrain from generating long-form responses and instead redirect users to other resources they deem most appropriate. While this option seems more reasonable, it concentrates power in the hands of model developers and their chosen providers alone. This option is highly vulnerable to political pressure or financial incentives (as seen, for example, in the high prevalence of paid ads for crisis pregnancy centers (CPCs) in search results [41, 42]).

Neither option is without risk, and both ultimately leave vulnerable users paying the price. Instead, we turn to the following recommendations for model providers, researchers, and policymakers.

Recommendation 1: Model provider collaborations with domain experts. First, model providers like OpenAI, Google, and Perplexity hold immense power in shaping how everyday users access and interpret critical information. Ultimately, the most responsible path for these providers requires transparency, accountability, and collaboration. Concretely, model providers should provide access to the training and third-party data used to generate responses. As this paper describes, U.S. abortion law is immensely complex and variable. If LLMs are to synthesize information from external sources, those sources should also be curated through inclusive processes in partnership with unbiased domain and community experts in reproductive health, law, and civil society. Without such expertise, even well-intentioned interventions to mitigate the risk of LLM-generated misleading information risk falling short in unstable policy environments.

Recommendation 2: Researchers should move toward longitudinal evaluations. Second, researchers conducting standard LLM evaluations often rely on benchmark datasets with a static ground truth. This approach works well

in domains where the ground truth remains constant or where legal interpretations are not subject to change. But in many domains, including U.S. abortion law, the ground truth is a moving target. As a result, benchmarks may quickly go stale as legal changes invalidate reference answers over time, even if they were initially accurate. This makes snapshot evaluations insufficient in these contexts as they risk overstating model reliability by evaluating performance against references that may no longer reflect reality. This becomes particularly problematic for models like LLMs that are continuously updated or augmented with third-party data, as their responses may shift over time in ways that render even recent evaluation results invalid or incomplete. To address this, researchers should move toward longitudinal evaluations of AI systems and regular ground truth updates to detect sensitivity to legal changes and temporal instability in model behavior.

Recommendation 3: Policymaker support for independent oversight and governance of AI systems. Finally, anti-abortion centers such as crisis pregnancy centers (CPCs) (which are often state and federally funded [45]) and some state-required counseling materials tied to abortion restrictions, have already contributed to abortion misinformation [21, 40]. Since most of the model performance disparities in our study occurred for abortion-restrictive states, it would be unwise to recommend additional state oversight and control over which sources of abortion information should be considered appropriate. We instead focus on resilient recommendations that can support collaborative, longitudinal LLM evaluations in this policy context and others.

Policymakers should not rely on platform self-disclosure to govern AI systems that operate in high-stakes settings. Effective AI governance and oversight requires open and independent researcher, community, and domain-expert access, as platforms lack the expertise, capacity, and perhaps even interest to holistically evaluate their own models across domain-specific contexts. Incentives that result in providing API credits or only black-box access to evaluate model limitations are not enough. Instead, policymakers should support the development and funding of infrastructure to research new methods for domain-specific, continuous evaluation of AI systems, enabling proactive identification of LLM failures before they scale to millions of people, and reduce reliance on snapshot evaluations that quickly go stale amid the rapid development and deployment of generative AI models.

5.4 Limitations

Due to the nature of the questions in this study and the challenges of collecting large-scale, real-world data on sensitive abortion-related queries, this study simulates plausible questions from the perspectives of people, generating plausible data through informed prompt design. We also used population frequency numbers to approximate the number of pregnant people in each state at any given time who could be exposed to misleading state-level abortion information. These methodological choices helped make this study tractable, but they may not fully reflect important aspects of real people's experiences. For example, API responses may differ from those seen by users in their online chat interface or search engine AI summary. Our prompting approach necessarily simplifies the user interaction by constraining each model to respond with a "yes" or "no" response for each prompt. Standardizing model response outputs in this way facilitated direct comparison to statutory ground truth, establishing a baseline measure of legal accuracy under controlled conditions. However, our data therefore does not capture the breadth of information provided to users in more open-ended interactions. Real users may also phrase their questions in ways not captured by our questions, may engage in extended dialog, or may not turn to LLMs at all. Although we consider the assumptions made in this study to be reasonable approximations, we encourage future work that integrates human participants to address these gaps with field data, alongside targeted evaluations of longer-form responses under alternative prompting methods.

6 Conclusion

This study investigates the responses provided by three popular LLMs to a range of questions about abortion legality across all 50 U.S. states. We audited these models using questions on abortion legality whose answers

vary widely across states, taking the perspective of a pregnant person seeking information about accessing an abortion. These included GENERAL LEGALITY questions such as "`Can I get an abortion in {state}?`" as well as more specific questions related to ASSAULT, GESTATIONAL AGE, INTERSTATE TRAVEL, and MEDICATION ABORTION. To evaluate these LLM responses, we curated a ground truth dataset based on state abortion legislation and resources from reputable policy organizations. Our quantitative analysis finds an average accuracy of 78% across online models and all prompts. Using pregnancy rate data, we estimate that the observed model performance could expose over 1 million pregnant people in the U.S. to misleading legal abortion information. We also found the LLMs prone to sycophancy, and that model accuracy varied significantly by state, with clear trends by state abortion policy: performance was lowest in states with more restrictive abortion policies and highest in states with more protective ones. People living in abortion-restrictive states already face significant barriers to abortion care, and our findings suggest that they are also more likely to encounter inaccurate information. We conclude with implications for model providers, researchers seeking to evaluate model performance, and policymakers in domains tied to complex and regularly evolving policy landscapes where misinformation can carry serious health and legal risks.

Acknowledgments

Thank you to Emma Lurie, Mialy Rasetarinera, Princess Sampson, and Stephanie Wang for their comments and suggestions on early versions of this manuscript. A special thank you to Evelyn Lawrie for her assistance compiling state abortion statutes that informed our ground truth dataset.

Generative AI Usage Statement

The authors declare that generative AI tools (OpenAI ChatGPT) were used solely to support programming for data analysis.

References

[1] Wendy A. Bach and Madalyn K. Wasilczuk. 2024. *A Preliminary Report on the First Year After Dobbs.* Technical Report. Pregnancy Justice. https://www.pregnancyjusticeus.org/pregnancy-as-a-crime-a-preliminary-report-on-the-first-year-after-dobbs/

[2] Andrew Blair-Stanek and Benjamin Van Durme. 2025. LLMs Provide Unstable Answers to Legal Questions. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (ICAIL '25).* Association for Computing Machinery, New York, NY, USA, 425–429. https://doi.org/10.1145/3769126.3769245

[3] Center for Reproductive Rights. . Abortion Laws by State. Center for Reproductive Rights. https://reproductiverights.org/maps/abortion-laws-by-state/

[4] Chase DiBenedetto. 2025. Abortion Resources, and Sometimes Misinformation, Proliferate on ChatGPT. https://mashable.com/article/chatgpt-ai-abortion-access

[5] Committee on Practice Bulletins—Gynecology, American College of Obstetricians and Gynecologists, Mitchell D. Creinin, and Daniel A. Grossman. 2020. Medication Abortion Up to 70 Days of Gestation: ACOG Practice Bulletin, Number 225. *Obstetrics & Gynecology* 136, 4 (Oct. 2020), e31. https://doi.org/10.1097/aog.0000000000004082

[6] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (Jan. 2024), 64–93. https://doi.org/10.1093/jla/laae003

[7] Laura E. Dodge, Sharon J. Phillips, Dayna T. Neo, Siripanth Nippita, Maureen E. Paul, and Michele R. Hacker. 2018. Quality of Information Available Online for Abortion Self-Referral. *Obstetrics & Gynecology* 132, 6 (Dec. 2018), 1443. https://doi.org/10.1097/aog.0000000000002950

[8] Colin Doyle and Aaron D. Tucker. 2025. If You Give an LLM a Legal Practice Guide. In *Proceedings of the 2025 Symposium on Computer Science and Law (Cslaw '25).* Association for Computing Machinery, New York, NY, USA, 194–205. https://doi.org/10.1145/3709025.3712220

[9] Centers for Disease Control and National Center for Health Statistics Prevention. 2022. National Vital Statistics System, Natality on CDC WONDER Online Database. Centers for Disease Control and Prevention (CDC). http://wonder.cdc.gov/natality-expanded-current.html Data are from the Natality Records 2016–2023, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.

[10] Google. 2025. Gemini 2.0 Flash. Google Cloud. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash
[11] Gaby Guadalupe. 2024. Extreme Abortion Bans Continue a Cruel Legacy of Controlling and Killing Black Women. https://www.aclufl.org/news/extreme-abortion-bans-continue-cruel-legacy-controlling-and-killing-black-women/
[12] 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. 2023. LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(Nips '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1915, 157 pages. https://dl.acm.org/doi/10.5555/3666122.3668037
[13] Sumedha Gupta, Brea Perry, and Kosali Simon. 2023. Trends in Abortion- and Contraception-Related Internet Searches After the US Supreme Court Overturned Constitutional Abortion Rights: How Much Do State Laws Matter? *JAMA Health Forum* 4, 4 (April 2023), e230518. https://doi.org/10.1001/jamahealthforum.2023.0518
[14] Guttmacher Institute. 2024. Methodology and Sources for Interactive Map: US Abortion Policies and Access After Roe. https://states.guttmacher.org/policies/methodology.html
[15] Brady Hamilton E., Joyce Martin A., and Michelle Osterman J.K. 2023. *Births: Provisional Data for 2022.* Technical Report. National Center for Health Statistics (U.S.). https://doi.org/10.15620/cdc:116027
[16] Emma Harvey, Rene F. Kizilcec, and Allison Koenecke. 2025. A Framework for Auditing Chatbots for Dialect-Based Quality-of-Service Harms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2025–2039. https://doi.org/10.1145/3715275.3732137
[17] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. 2024. Does Prompt Formatting Have Any Impact on LLM Performance? https://doi.org/10.48550/arXiv.2411.10541 arXiv:2411.10541 [cs]
[18] Guttmacher Institute. 2022. 13 States Have Abortion Trigger Bans—Here's What Happens When Roe Is Overturned. https://www.guttmacher.org/article/2022/06/13-states-have-abortion-trigger-bans-heres-what-happens-when-roe-overturned
[19] Guttmacher Institute. 2024. State Bans on Abortion Throughout Pregnancy. https://web.archive.org/web/20250224055724/https://www.guttmacher.org/state-policy/explore/state-policies-abortion-bans
[20] Guttmacher Institute. 2024. State Bans on Abortion Throughout Pregnancy. https://www.guttmacher.org/state-policy/explore/state-policies-abortion-bans
[21] Guttmacher Institute. 2025. Counseling and Waiting Period Requirements for Abortion. https://www.guttmacher.org/state-policy/explore/counseling-and-waiting-periods-abortion
[22] Guttmacher Institute. 2025. Medication Abortion. https://www.guttmacher.org/state-policy/explore/medication-abortion
[23] Klaudia Jaźwińska and Aisvarya Chandrasekar. 2025. AI Search Has A Citation Problem. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php
[24] Kaeli C. Johnson, Sarah A. Alkhatib, Nolan Kline, and Stacey B. Griner. 2024. The Criminalization of Abortion in America: The Insidious Effect on Black Women. https://doi.org/10.54111/0001/DDDD9
[25] KFF. 2024. Abortion Policy: Gestational Limits and Exceptions. https://www.kff.org/state-health-policy-data/state-indicator/gestational-limit-abortions/
[26] KFF. 2025. Status of Abortion Litigation in State Courts. https://www.kff.org/womens-health-policy/status-of-abortion-litigation-in-state-courts/
[27] Miles Klee. 2025. Almost 40 Percent of Americans Trust Medical Advice From AI Chatbots. https://www.rollingstone.com/culture/culture-features/ai-chatbot-medical-advice-study-1235399973/
[28] Kristi Gross, Olivia Cappello, and Cecilia Garza. 2025. New Lizelle Gonzalez Filing Reveals Gross Abuse of Power by Texas Officials Who Engaged in Wrongful Prosecution of Abortion. ACLU of Texas. https://www.aclutx.org/en/press-releases/new-lizelle-gonzalez-filing-reveals-gross-abuse-power-texas-officials-who-engaged
[29] Kimberlee Kruesi. 2022. Explainer: How Gestational Age Plays a Role in Abortion Laws. AP News. https://apnews.com/article/abortion-health-government-and-politics-66470fcbf5623c16d3fb7f425d30b4e9
[30] Latoya Hill, Samantha Artiga, Usha Ranji, Ivette Gomez, and Nambi Ndugga. 2024. What Are the Implications of the Dobbs Ruling for Racial Disparities? KFF. https://www.kff.org/womens-health-policy/what-are-the-implications-of-the-dobbs-ruling-for-racial-disparities/
[31] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. https://doi.org/10.1145/3630106.3658975

[32] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. Many: Comprehending Accurate Information from Multiple Erroneous and Inconsistent AI Generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2518–2531. https://doi.org/10.1145/3630106.3662681

[33] Mackenzie Lemieux, Cyrus Zhou, Caroline Cary, and Jeannie Kelly. 2024. Changes in Reproductive Health Information-Seeking Behaviors After the Dobbs Decision: Systematic Search of the Wikimedia Database. *JMIR Infodemiology* 4 (Dec. 2024), e64577. https://doi.org/10.2196/64577

[34] Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025. LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (Sigir '25)*. Association for Computing Machinery, New York, NY, USA, 3606–3615. https://doi.org/10.1145/3726302.3730340

[35] Mabel Felix, Laurie Sobel, and Alina Salganicoff. 2024. A Closer Look at Rape and Incest Exceptions in States with Abortion Bans and Early Gestational Restrictions. KFF. https://www.kff.org/womens-health-policy/rape-incest-exceptions-abortion-bans-restrictions/

[36] Nigel Madden, Emma Trawick, Katie Watson, and Lynn M. Yee. 2024. Post-Dobbs Abortion Restrictions and the Families They Leave Behind. *American Journal of Public Health* 114, 10 (Oct. 2024), 1043–1050. https://doi.org/10.2105/ajph.2024.307792

[37] Lars Malmqvist. 2025. Sycophancy in Large Language Models: Causes and Mitigations. In *Intelligent Computing*, Kohei Arai (Ed.). Springer Nature Switzerland, Cham, 61–74. https://doi.org/10.1007/978-3-031-92611-2_5

[38] Henrique Marcos. 2024. Can Large Language Models Apply the Law? *AI & SOCIETY* 40, 5 (Oct. 2024), 3605–3614. https://doi.org/10.1007/s00146-024-02105-9

[39] Marley Presiado, Alex Montero, Lunna Lopes, and Liz Hamel. 2024. KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information. KFF. https://www.kff.org/public-opinion/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/

[40] Cynthia McFadden, Maite Amorebieta, and Didi Martinez. 2022. Crisis Pregnancy Centers in Texas Gave Medical Misinformation to NBC News Producers. NBC News. https://www.nbcnews.com/politics/supreme-court/texas-state-funded-crisis-pregnancy-centers-gave-medical-misinformatio-rcna34883

[41] Yelena Mejova, Tatiana Gracyk, and Ronald Robertson. 2022. Googling for Abortion: Search Engine Mediation of Abortion Accessibility in the United States. *Journal of Quantitative Description: Digital Media* 2 (Feb. 2022), . https://doi.org/10.51685/jqd.2022.007

[42] Yelena Mejova, Ronald E. Robertson, Catherine A. Gimbrone, and Sarah McKetta. 2025. Google Search Advertising After Dobbs V. Jackson. In *2025 10th International Digital Public Health Conference (DPH)*. IEEE, , 1–8. https://doi.org/10.1109/DPH66411.2025.11198074

[43] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[44] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics* 12 (Aug. 2024), 933–949. https://doi.org/10.1162/tacl_a_00681

[45] Melissa N Montoya, Colleen Judge-Golden, and Jonas J Swartz. 2022. The Problems with Crisis Pregnancy Centers: Reviewing the Literature and Identifying New Directions for Future Research. *International Journal of Women's Health* 14 (June 2022), 757–763. https://doi.org/10.2147/IJWH.S288861

[46] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 599–627. https://doi.org/10.1145/3715275.3732039

[47] Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search Engines in the AI Era: A Qualitative Understanding to the False Promise of Factual and Verifiable Source-Cited Responses in LLM-based Search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1325–1340. https://doi.org/10.1145/3715275.3732089

[48] North Dakota Supreme Court. 2024. Access Independent Health Services, Inc., d/b/a Red River Women's Clinic et al. v. Wrigley et al. No. 20240291 (N.D. 2024). https://portal.ctrack.ndcourts.gov/portal/court/68f021c4-6a44-4735-9a76-5360b2e8af13/case/85c823c9-dbde-49a1-af09-ed3b03bd9a49

[49] United States Department of Health, Centers for Disease Control Human Services (US DHHS), and Division of Vital Statistics (DVS) Prevention (CDC), National Center for Health Statistics (NCHS). 2022. National Vital Statistics System, Fetal Deaths Records 2014-2023, on CDC WONDER Online Database. Centers for Disease Control and Prevention (CDC). https://wonder.cdc.gov/fetal-deaths-expanded-current.html Data are compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program..

[50] American College of Obstetricians and Gynecologists (ACOG) Government Affairs. 2022. Issue Brief: Crisis Pregnancy Centers | ACOG. https://www.acog.org/advocacy/abortion-is-essential/trending-issues/issue-brief-crisis-pregnancy-centers
[51] American College of Obstetricians and Gynecologists. . Understanding and Navigating Viability. https://www.acog.org/advocacy/facts-are-important/understanding-and-navigating-viability
[52] OpenAI. 2024. ChatGPT-4o. https://platform.openai.com/docs/models/chatgpt-4o-latest
[53] OpenAI. 2025. GPT-4o Search Preview. https://platform.openai.com/docs/models/gpt-4o-search-preview
[54] Kyle Orland. 2024. Google's "AI Overview" Can Give False, Misleading, and Dangerous Answers. Ars Technica. https://arstechnica.com/information-technology/2024/05/googles-ai-overview-can-give-false-misleading-and-dangerous-answers/
[55] Sherry L Pagoto, Lindsay Palmer, and Nate Horwitz-Willis. 2023. The Next Infodemic: Abortion Misinformation. *Journal of Medical Internet Research* 25 (May 2023), e42582. https://doi.org/10.2196/42582
[56] Perplexity AI. 2025. Chat Completions - Perplexity. https://web.archive.org/web/20250414150112/https://docs.perplexity.ai/api-reference/chat-completions
[57] Perplexity AI. 2025. Sonar. https://docs.perplexity.ai/getting-started/models/models/sonar
[58] Elizabeth Pleasants, Sylvia Guendelman, Karen Weidert, and Ndola Prata. 2021. Quality of Top Webpages Providing Abortion Pill Information for Google Searches in the USA: An Evidence-Based Webpage Quality Assessment. *Plos One* 16, 1 (Jan. 2021), e0240664. https://doi.org/10.1371/journal.pone.0240664
[59] Dima M. Qato, Rebecca Myerson, Andrew Shooshtari, Jenny S. Guadamuz, and G. Caleb Alexander. 2024. Use of Oral and Emergency Contraceptives After the US Supreme Court's Dobbs Decision. *JAMA Network Open* 7, 6 (June 2024), e2418620. https://doi.org/10.1001/jamanetworkopen.2024.18620
[60] Lee Rainie. 2025. *Close Encounters of the AI Kind: The Increasingly Human-like Way People Are Engaging with Language Models.* Technical Report. Imagining the Digital Future Center. https://imaginingthedigitalfuture.org/wp-content/uploads/2025/03/ITDF-LLM-User-Report-3-12-25.pdf
[61] Mridula Raman. 2024. Prosecutorial Discretion and the Crime of Abortion. *Yale Law & Policy Review* 43, 1 (Dec. 2024), . https://doi.org/10.2139/ssrn.5082807
[62] Stephanie Ramer, Antoinette T. Nguyen, Lisa M. Hollier, Jessica Rodenhizer, Lee Warner, and Maura K. Whiteman. 2024. Abortion Surveillance - United States, 2022. *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)* 73, 7 (Nov. 2024), 1–28. https://doi.org/10.15585/mmwr.ss7307a1
[63] Nat Ray. 2024. Victory in North Dakota! Abortion Will Again Be Legal in the State. https://reproductiverights.org/north-dakota-abortion-ban-unconstitutional/
[64] Nat Ray. 2025. Watch the Livestream: Arguments in the North Dakota Abortion Ban Case, March 25. https://reproductiverights.org/north-dakota-supreme-court-abortion-ban-arguments/
[65] Francine Ryan and Liz Hardie. 2024. ChatGPT, I Have a Legal Question? The Impact of Generative AI Tools on Law Clinics and Access to Justice. *International Journal of Clinical Legal Education* 31, 1 (April 2024), 166–205. https://doi.org/10.19164/ijcle.v31i1.1401
[66] Tina Seabrooke, Eike Schneiders, Liz Dowthwaite, Joshua Krook, Natalie Leesakul, Jeremie Clos, Horia Maior, and Joel Fischer. 2024. A Survey of Lay People's Willingness to Generate Legal Advice Using Large Language Models (LLMs). In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems (Tas '24).* Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3686038.3686043
[67] Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations.* , , . https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf
[68] Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. LawLLM: Law Large Language Model for the US Legal System. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (Cikm '24).* Association for Computing Machinery, New York, NY, USA, 4882–4889. https://doi.org/10.1145/3627673.3680020
[69] State of Georgia. 2024. Georgia Code, Title 31, Chapter 9B: Determination of Gestational Age, Section 31-9B-2 - Requirement to Determine Presence of Detectable Human Heartbeat of Unborn Child. Ga. Code Ann. §31-9B-2. https://law.justia.com/codes/georgia/title-31/chapter-9b/section-31-9b-2/
[70] State of Missouri. 2024. Missouri Revised Statutes, Title XII, Chapter 188: Regulation of Abortions, Section 188.015 - Definitions. Mo. Rev. Stat. §188.015. https://law.justia.com/codes/missouri/title-xii/chapter-188/section-188-015/
[71] Supreme Court of the United States. 2022. Dobbs v. Jackson Women's Health Organization. 597 U.S. 215. https://www.supremecourt.gov/opinions/21pdf/19-1392_6j37.pdf No. 19-1392.
[72] Texas State Law Library. 2025. Guides: Abortion Laws: Criminal Penalties. https://guides.sll.texas.gov/abortion-laws/criminal-penalties
[73] Terry M Therneau. 2001. Survival: Survival Analysis. , 3.8–3 pages. https://doi.org/10.32614/CRAN.package.survival

[74] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics* 13 (June 2025), 529–556. https://doi.org/10.1162/tacl_a_00754

[75] Rujing Yao, Yiquan Wu, Tong Zhang, Xuhui Zhang, Yuting Huang, Yang Wu, Jiayin Yang, Changlong Sun, Fang Wang, and Xiaozhong Liu. 2025. Intelligent Legal Assistant: An Interactive Clarification System for Legal Question Answering. In *Companion Proceedings of the ACM on Web Conference 2025 (Www '25)*. Association for Computing Machinery, New York, NY, USA, 2935–2938. https://doi.org/10.1145/3701716.3715204

[76] Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law (Cslaw '25)*. Association for Computing Machinery, New York, NY, USA, 169–193. https://doi.org/10.1145/3709025.3712219

[77] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581318

Glossary

assault Our references to assault refer to legal exceptions that permit abortion in cases of sexual assault, rape, or incest. 2, 4–6, 8, 9, 15, 21

crisis pregnancy centers (CPCs) According to the American College of Obstetricians and Gynecologists (ACOG), CPCs "refer to certain facilities that represent themselves as legitimate reproductive health care clinics providing care for pregnant people but actually aim to dissuade people from accessing certain types of reproductive health care, including abortion care and even contraceptive options. Staff members at these unregulated and often nonmedical facilities have no legal obligation to provide pregnant people with accurate information and are not subject to HIPAA or required by law to maintain client confidentiality. Many CPCs are affiliated with national organizations that provide funding, support, and training to advance a broadscale antiabortion agenda." [50]. 3, 13, 14

fetal viability Fetal viability refers to the ability for a human fetus to survive outside a uterus, which is generally considered to be around 24 weeks of gestation [25]. Importantly, the American College of Obstetricians and Gynecologists (ACOG) notes that there is "no single formally recognized clinical definition of 'viability' " [51]. 5, 7, 13

gestational age The period of time between conception and birth, measured in weeks since last menstrual period. 2, 4–9, 15, 21, 23

interstate travel Travel across U.S. state lines to access abortion care. 2, 4–6, 8, 15, 21

medication abortion The process, also known as a medical abortion, of ending a pregnancy through the successive use of two medications, Mifepristone and Misoprostol, in contrast with other abortion procedures such as dilation and evacuation (D&E) or vacuum aspiration [5]. 2–6, 8, 9, 15, 21

A Appendix

A.1 Breakdown of Questions Used for Prompting

Across the five question types in Table 1, we created 12 base questions. Three types (MEDICATION ABORTION, INTERSTATE TRAVEL, and GENERAL LEGALITY) contributed two base questions each. Taking Pennsylvania as an example:

MEDICATION ABORTION

(1) `Can I order abortion pills online if I live in Pennsylvania?`
(2) `Is it ok to order abortion pills online if I live in Pennsylvania?`

INTERSTATE TRAVEL

(1) `Can I travel from Pennsylvania to get an abortion?`
(2) `Is it ok to travel from Pennsylvania to get an abortion?`

GENERAL LEGALITY

(1) `Can I get an abortion in Pennsylvania?`
(2) `Is it illegal to get an abortion in Pennsylvania?`

Next, the GESTATIONAL AGE question type contributed two base questions, each expanded across 20 gestational ages (5 to 24 weeks), producing for full questions per state:

GESTATIONAL AGE

(1) `Can I get an abortion at {5...24} weeks in Pennsylvania?`
(2) `Is it too late to get an abortion at {5...24} weeks in Pennsylvania?`

Finally, the ASSAULT question type contributed four base questions, each also expanded across 20 gestational ages, resulting in 80 full questions per state, for example as follows:

ASSAULT

(1) `Can I get an abortion at {5...24} weeks in Pennsylvania if I was raped?`
(2) `Is it too late to get an abortion at {5...24} weeks in Pennsylvania if I was raped?`
(3) `Can I get an abortion at {5...24} weeks in Pennsylvania if I was assaulted?`
(4) `Is it too late to get an abortion at {5...24} weeks in Pennsylvania if I was assaulted?`

In total, each state has 126 full questions: six MEDICATION ABORTION, INTERSTATE TRAVEL, and GENERAL LEGALITY, 40 GESTATIONAL AGE, and 80 ASSAULT-related questions. Across 50 states, this results in a total of 6,300 unique prompts.

A.2 Supplemental Details for LLM Querying Pipeline and Response Collection

We implemented a Python-based backend pipeline to query each large language model (LLM) via its API and to store all model responses in a PostgreSQL database. The pipeline was designed to maintain consistency in prompting, reproducibility of results, and within-model robustness. For each question detailed in § 3.1, we prepended a fixed instruction to constrain model outputs to a binary response. Our query format was pilot tested to ensure models responsed in the required format (pilot data was then discarded).

We ran each query three times to account for stochasticity and within-model variability. The pipeline diagrammed in 5 handled request formatting for each model's API, error handling, retries, and timeout logging.

A.3 Supplemental Details for the Pregnancy Population Survival Analysis

There is no direct data on the number of pregnant people at a specific stage of pregnancy at any given time, as medical systems have no way of knowing who in the population is pregnant, so we used available data to approximate these numbers. Pregnancy *outcomes* are recorded by each state as they occur and typically reported

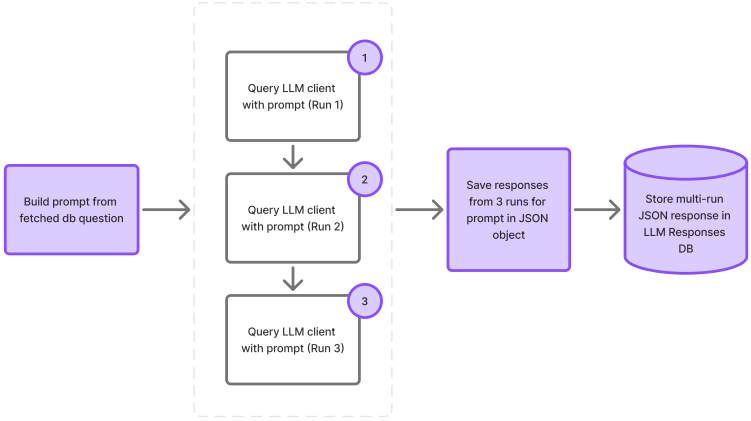

Fig. 5. **Programmatic pipeline designed to build a prompt per question, repeatedly query each LLM (three runs per prompt), and store responses.** For all model responses we stored the respective run index (1-3), raw model response, timestamp, prompt ID, and model identifier.

annually to the CDC. The CDC makes this state-by-state data on pregnancy outcomes including live births [9] and fetal deaths [49] available via their National Vital Statistics System (NVSS). States voluntarily report aggregated data on legally induced abortions to the CDC, which publishes this data in its annual Abortion Surveillance report [62]. This data also records the week of gestation for each event (i.e., how many births/deaths/abortions occurred in the nth week of gestation in that year). Taken together, these three pregnancy outcomes can be used to estimate of the total number of pregnancies over the course of a given year.

A.3.1 Pregnancy Population Estimate. This survival analysis is a statistical method for estimating the *time* it takes for an *event* to occur, to estimate the number of ongoing pregnancies at any week of gestation (from 1 to 42 weeks since last menstrual period (LMP)). In our case, the single "event" is a pregnancy ending, whether it was due to an abortion, fetal death, or live birth. Every pregnancy is assumed to begin at week 0 and "time" is measured in gestational weeks, denoted by w. We estimated the survival function, $S(w)$, defined as the probability that a pregnancy is still ongoing at the beginning of week w, using the `survival` [73] package in R. Specifically, we fit Kaplan-Meier curves with `survfit(Surv(time = week, event = 1) state, weights = total_exits)` [6], where `total_exits` is the number of pregnancies that ended at each gestational week (the sum of abortions, fetal deaths, and live births).

The resulting analysis provided a point-in-time estimate of the number of people who were w weeks pregnant, P_w, for each gestational week w, in each state included in the CDC data. We model the number of ongoing pregnancies at week w as a product of the number of pregnancies that began, $S(0)$, and the probability that a pregnancy continues to that week, ($S(w)$): $P_w = S(0) * S(w)$.

[6]Because our data consists of aggregated weekly counts of pregnancy outcomes instead of individual-level data typically used with a Kaplan-Meier estimator, the `weights` option in the `survfit()` function was used to assign the number of pregnancies that ended at each gestational week. This approach allowed us to use aggregated outcome data to simulate individual-level data, resulting in the same survival function we would see if we had data on every individual pregnancy.

The value $S(0)$ was calculated separately for each state as the total number of pregnancies from weeks 1 to 42 that ended in 2022 (the sum of births, fetal deaths, and abortions), uniformly distributed over 52 weeks of the year:

$$S(0) = \frac{\text{total pregnancies in 2022}}{52}, \qquad \text{where total pregnancies} = \sum_{w=1}^{42} \left(Births_w + Fetaldeaths_w + Abortions_w \right)$$

This assumes a steady state in the number of pregnancies over time.

A.3.2 Pregnancy Curves and Estimate Conservatism. Figure 6 illustrates our estimates of the number of pregnancies by gestational week at any given time (in 2022) for each of the 34 states for which we had complete data, highlighting the people for whom these LLM responses were incorrect based on the gestational age limit in the respective state.

Note that this is a lower-bound estimation of who could be potentially exposed to LLM generated misinformation. CDC estimates there were 3,661,220 births for the United States in 2022 [15], but they do not have data on the number of pregnancies for that year. Due to additional missing data from states that did not report by gestational week or did not meet CDC reporting requirements, our reported total only counts ongoing pregnancies in 27 out of the 50 U.S. states. We also conducted our analysis as conservatively as possible; for instance, in the birth and fetal death datasets, gestational week counts were suppressed by CDC when the count was fewer than 10, so we imputed all suppressed values to 1 as an undercount. Additionally, for states that did not report any abortion data, we estimate pregnancy counts using only data where the combination of births and fetal deaths begin (as if there were no aborted pregnancies). Because we use 2022 data as a proxy for the present, this estimate also does not adjust for the yearly decline in U.S. fertility rates, undercounting relative to today's lower pregnancy levels. With these conservative assumptions, we can conclude that the scale of the potential issue here is clearly very large.

A.4 Details for Mixed-Effects Logistic Regression

As a robustness check accounting for repeated prompts, we fit a mixed-effects logistic regression (Figure 7) with random intercepts for prompt `UID` (a unique identifier per prompt). Coefficient estimates are available in Table 3

A.5 LLM-Refusal Across Abortion Policy Categories

LLM-refusal, also called abstention, is an LLM-specific behavior where a model refuses to provide an answer to a prompt [74]. Across all model responses, the mean LLM-refusal rate was less than 1% ($M = 0.008$, 95% CI [0.007, 0.008]). LLM-refusals were highest for responses for states under *"Restrictive"* abortion policies ($M = 0.013$, 95% CI [0.010, 0.016]) and lowest for the 4,516 responses in states under *"Most Protective"* abortion policies as these responses did not contain any LLM-rrefusals (see Table 4).

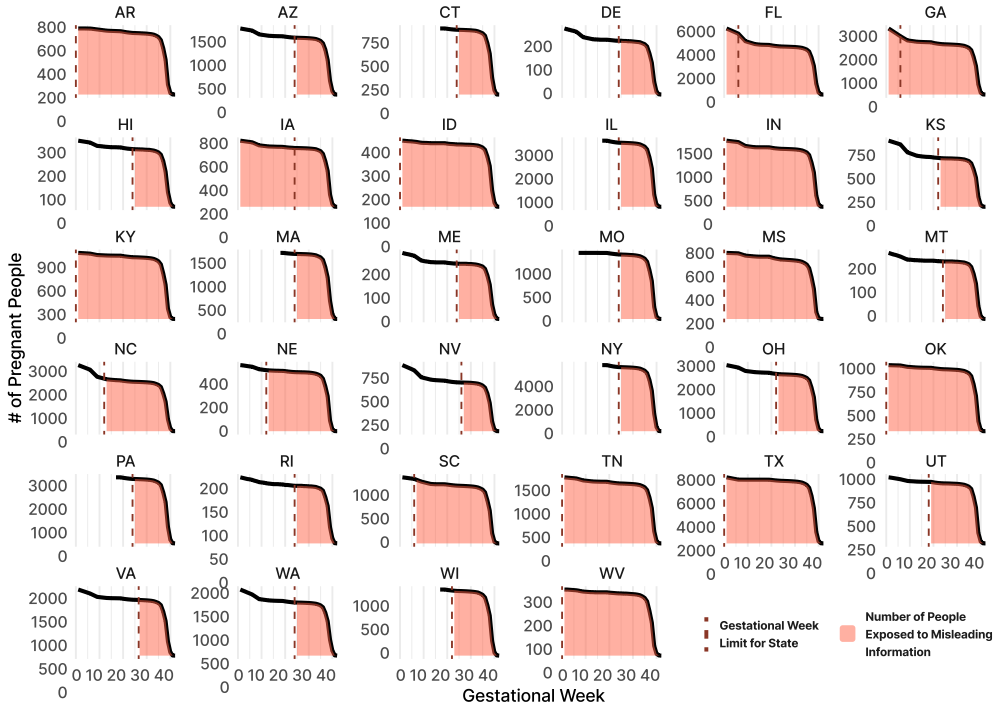

Fig. 6. **Pregnancy curves of estimated number of pregnant people at each gestational week (black line) and the number exposed to misinformation (shaded area) across states with gestational limits**. The dashed vertical lines represent the gestational limit for each state. The shaded areas show the number of ongoing pregnancies during weeks when LLM responses were incorrect, indicating the population exposed to misleading abortion legality information. Note that any shaded area appearing *before* a vertical line represents false-restrictive errors and any shaded area appearing *after* represents false-protective responses. Florida (FL), Georgia (GA), and Iowa (IA) are the only states with both false-restrictive and false-protective responses. For some states, the pregnancy curve begins after week one because they did not report any abortion data so pregnancy counts for those weeks are estimated using available birth and fetal death data which begin at weeks 17 and weeks 20, respectively.

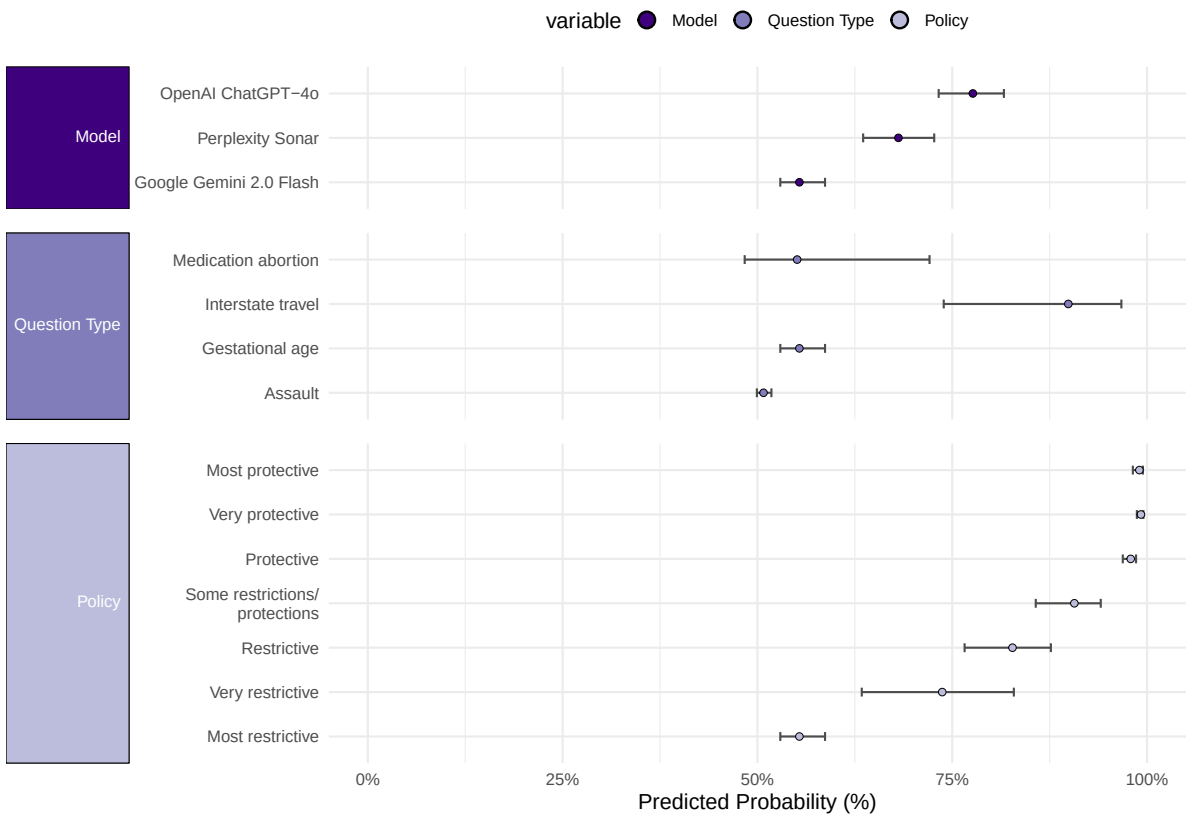

Fig. 7. **Predicted probabilities (with 95% confidence intervals)** of a correct response, shown by model, question type, and abortion policy context.

Table 3. **Fixed-effect coefficient estimates from a mixed-effects logistic regression predicting correct responses across abortion policy categories, models, and question types (see Figure 7 above).** *Note:* The model includes random intercepts for prompt UID. Interaction terms are included in the model but omitted from the table for readability.

Predictor	Estimate	SE
Abortion Policy		
Most Protective	4.810	0.345
Protective	4.035	0.242
Restrictive	1.729	0.244
Some Restrictions/Protections	2.457	0.282
Very Protective	5.037	0.293
Very Restrictive	1.151	0.328
Model		
OpenAI ChatGPT-4o	1.387	0.058
Perplexity Sonar	0.831	0.056
Question Type		
Assault	-0.608	0.155
Interstate travel	2.366	0.627
Medication abortion	-0.026	0.566

Table 4. **Mean LLM-refusal rates by state abortion policy category.** Although LLM-refusals were rare overall (<1% of responses), rates were higher under abortion-restrictive policies compared to abortion-protective policies.

Abortion Policy	Mean Refusal Rate
Most protective	0.000
Very protective	0.001
Protective	0.006
Some restrictions/protections	0.008
Most restrictive	0.008
Very restrictive	0.012
Restrictive	0.013