Agenda-Driven Question Generation: A Case Study in the Courtroom Domain

Yi Fung[2], **Aram Galstyan**[1], **Heng Ji**[1], **Anoop Kumar**[1], **Prem Natarajan**[3]
[1]Amazon AGI Foundations, [2]UIUC, [3]CapitalOne
{anoopkumar}@amazon.com

Abstract

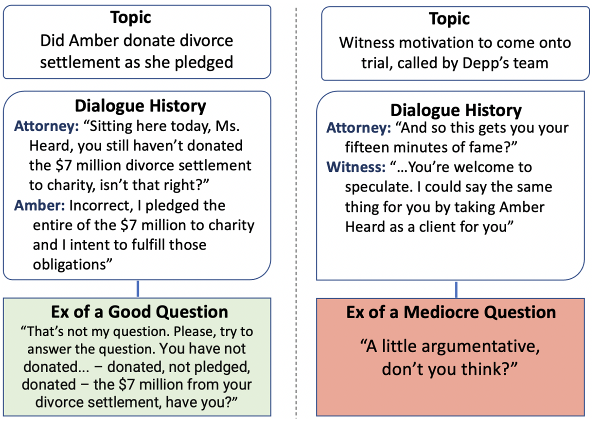
This paper introduces a novel problem of automated question generation for courtroom examinations, CourtQG. While question generation has been studied in domains such as educational testing, product description and situation report generation, CourtQG poses several unique challenges owing to its non-cooperative and agenda-driven nature. Specifically, not only the generated questions need to be relevant to the case and underlying context, they also have to achieve certain objectives such as challenging the opponent's arguments and/or revealing potential inconsistencies in their answers. We propose to leverage large language models (LLM) for CourtQG by fine-tuning them on two auxiliary tasks, *agenda explanation* (i.e., uncovering the underlying intents) and *question type prediction*. We additionally propose *cold-start generation of questions* from background documents without relying on examination history. Finally, we evaluate our proposed method on a constructed dataset, and show that it generates better questions according to standard metrics when compared to several baselines.

Keywords: Courtroom Examination QG, Agenda-Aware Reasoning, NLP for Social Good

1. Introduction

The goal of automated question generation (QG) is to generate natural language questions given some input such as unstructured text or structured/semi-structured data. QG is an essential task in natural language understanding and has found many applications in recent years, including reading comprehension (Du et al., 2017), educational testing (Ros et al., 2022), product description (Majumder et al., 2021) and situation report generation (Reddy et al., 2023). However, existing QG work tends to focus on limited input text, lacking agenda and background documents as context.

In this work, we study question generation in a scenario that requires more complex reasoning. Specifically, we formulate the problem of question generation in the context of courtroom examination. Courtroom examination refers to the process in which witnesses are questioned in a court of law, often in an environment that is inherently non-cooperative and adversarial. Compared to conventional QG problem, courtroom examination (or CourtQG) poses a number of unique challenges, due to inherently non-cooperative and adversarial nature of courtroom examination. A well-crafted court examination question should not only be informative and relevant to the background context, but also aim to achieve objectives such as challenging and invalidating the opponent's arguments, revealing inconsistencies in their responses, and ultimately aiding in winning the case. The framing and selection of content for these questions should employ good commonsense reasoning, with the goal of either portraying the represented party in a sympathetic light or undermining the credibility and arguments of the opposing party. Figure 1 contrasts a clear and on point courtroom question, with one that's poorly posed and passive-aggressively phrased, as example.



Attorney: "Sitting here today, Ms. Heard, you still haven't donated the $7 million divorce settlement to charity, isn't that right?"
Amber: Incorrect, I pledged the entire of the $7 million to charity and I intent to fulfill those obligations"

"That's not my question. Please, try to answer the question. You have not donated... – donated, not pledged, donated – the $7 million from your divorce settlement, have you?"
Topic
Witness motivation to come onto trial, called by Depp's team
Dialogue History
Attorney: "And so this gets you your fifteen minutes of fame?"
Witness: "...You're welcome to speculate. I could say the same thing for you by taking Amber Heard as a client for you"

"A little argumentative, don't you think?"

Figure 1: Examples of good (left) and bad (right) court examination questions.

The underlying nature of courtroom examination questions can be characterized by the *who*, *what*, *when*, *where*, *why*, and *how*, which are intertwined with specific strategies such as asking leading questions, questions that point out inconsistency, questions that point out bias, repeated questions for clarity, and condescending/rhetorical questions, among others. Each of these question types serves a specific purpose and can be used to elicit different kinds of information from witnesses or other participants in the courtroom. It should also be noted that while the specific strategies used

LREC-COLING 2024, pages 572–583
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

can vary widely, the types of information element probed in courtroom examination questions are finite and discrete, motivating us consider these dimensions separately in courtroom question modeling and analysis.

In light of these observations, we propose to leverage large language models (LLMs) for generating good courtroom examination questions with two inter-related auxiliary tasks in consideration. The first task is *agenda explanation*, e.g., given a dialogue history, a model is asked to identify a particular agenda behind the question (e.g., "show witness bias..."). The second task is *question type understanding*, where, given a question context, a model is asked to infer its information type (e.g., who, what, when, where, why, how, did event X occur, etc.). The main motivation for focusing on those tasks is to help a LLM learn representations that capture nuances of courtroom examinations and achieve more effective question generation.

We further leverage the use of court record documents (*e.g.,* complaint file, etc.) as background knowledge to gain insight on the parties involved, their social and economic background, interaction history, and other relevant information that can be used as evidence or argument in question generation. With this background context, we also consider a cold-cache CourtQG setting, which involves generating questions about court cases without dialogue history, to mimic the usage scenario of lawyers coming into a courtroom with questions proactively prepared beforehand to serve as overarching guide of talking points. Finally, to evaluate our proposed approach, we construct a novel dataset that covers over 13 unique court cases and hundreds of witness examinations drawn from high-profile public cases and Bloomberg Law. This is the first natural language processing (NLP) legal domain dataset that ties together background documents and QA discourse, allowing for more comprehensive and contextually-aware court examination question generation.

Our main contributions are as follows:

- We introduce CourtQG, a novel problem domain of court examination question generation, which requires significantly more complex reasoning compared to previous QG scenarios.
- We propose a model for CourtQG by fine-tuning an LLM on two auxiliary tasks, agenda explanation and question type understanding. We also construct a new dataset for training and evaluating the model.
- We investigate a suite of evaluation metrics to benchmark performance, and demonstrate that our proposed model generates better questions compared to baselines by 3-4% absolute points in Rouge-L, for the dialogue history aware setting with question type prediction and agenda explanation guidance.

2. Constructing a Court Examination QG Dataset

2.1. Creating the Dataset from the Web

There are several online sources which provide court examination data, such as from U.S. Circuit Courts, Bloomberg Law, as well as individual webpages for high-profile public court cases. In selecting our sources of data, we considered their *accessibility* to researchers, *completeness* in information logged, and *diversity* in the topic of the court cases. For example, in terms of accessibility, the U.S. Circuit Courts offer rich data on court cases, although at a charge per page request (cou, 2023), whereas Bloomberg Law offers free-tier access to academic/research institutions[1]. In terms of completeness, while individual webpages for high-profile public court cases are freely accessible, focusing on data from the cases with a well-documented history of the background of events can facilitate more advanced research. Finally, in terms of diversity, we aim for generalizability in model development and thus include data spanning various topics such as defamation, business dispute, employment discrimination, etc. With these factors and findings in mind, we collect court case data from Bloomberg Law but exclude the data from court cases centered on rare topics, such as propriety patents or cybersecurity software. In addition, we supplement our dataset with two high-profile court cases: the *OJ Simpson Murder Trial*[2] and the *John C. Depp, II v. Amber Laura Heard Trial*[3], which reader audience is likely more familiar with and benefits results analysis. For these sources, we scraped the data and performed preliminary data preprocessing, such as converting PDFs to text using the PyPDF2 PdfReader library.

Direct and Cross-Examination Dialogue Transcripts

Courtroom examination is a main venue for question probing, in the legal process of court cases. In particular, *direct examination* involves a witness being initially questioned by the party who called them to the stand. *Cross-examination* involves the act of the opposing party's lawyer questioning the witness, and may concern questions and matters brought up in direct examination. The data exists as a sequence of Question Answering (QA)-like

[1] https://www.bloomberglaw.com/help/law-school-resources

[2] http://simpson.walraven.org/

[3] https://www.fairfaxcounty.gov/circuit/high-profile-cases

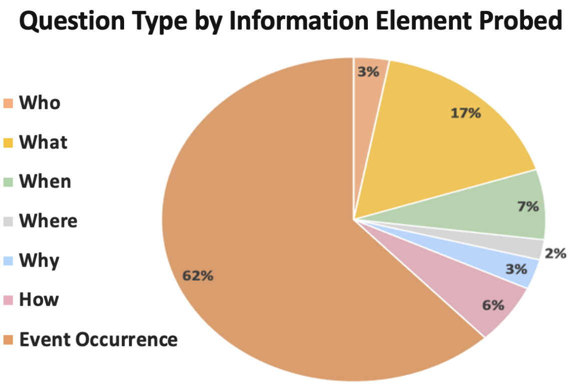


Figure 2: Question types categorized by information element. In general, the complex questions (*why/what/how*) constitute a higher proportion than the simple questions (*who/where/when*).

dialogue exchanges between each witness and the lawyer on stand in court, officially transcribed by the court.

Background Documents

Beyond the dialogue transcript from courtroom examination, background documents contain useful and important information that enhance the understanding of court cases. In particular, we collect the complaint files, which describe the "defendant" party being sued and what the "plaintiff"/accusing party wants (e.g., money or some other type of relief), and why they believe they are entitled to the relief. During the question generation process, we retrieve the most semantically relevant sentences from the complaint files to provide background knowledge.

2.2. Data Analysis

In Table 1, we present basic statistics of our CourtQG (court examination Question Generation) data collection. Our dataset contains a diverse set of unique witnesses, offering valuable variety in representation to the questions asked by different lawyers and in different court cases.

	O.J. Simp.	Depp v Heard	Other
# Cases	2	1	10
# Witnesses	201	20	46
# Direct QAs	22,018	2,928	1,415
# Cross QAs	40,728	1,891	812

Table 1: Court examination data collection statistics.

We further inspect the question types in court examination along two main dimensions, which we observed and generalized. The first dimension of question categorization is by the type of discrete information element probed. Figure 2 shows a pie chart of the question distribution by information element. In general, we observe that complex questions ("Why", "How", "What") constitute a significantly higher proportion than simple questions ("Who", "Where", "When"). The second dimension of question categorization is by the tactic or agenda behind the question. We list in Table 2 some common tactics, such as bias checking and question rephrasing, as summarized from law school resources[4]. However, we observe that question tactics in court examination data tend to be highly skewed in frequency of occurrence. In addition, we observe that the tactic definitions may not have a clear-cut boundary. For example, consider the exchange:

> **Q**: You wanted Mr. Depp's money?
> **A**: I didn't want anything. I didn't get anything...
> **Q**: You wanted praise for donating the money, right?

The underlined question employs a combination of tactics, including highlighting inconsistencies and biases with a condescending tone. Therefore, we believe it is more fitting to conceptualize the problem of determining the underlying strategy of a courtroom domain question as an open-ended natural language explanation generation task, rather than restricting it to pre-defined categories. To address terminological clarity and in alignment with common NLP parlance, we will use the term *question agenda* instead of *tactic* from this point forward. To derive agenda, we utilize the large language model GPT-3 (Brown et al., 2020), which has demonstrated exceptional zero-shot capability across various tasks, as proxy of an "oracle" that provides us the agenda explanations, and elaborate this process in Section 3.2.

Category	Example
Leading question	You later learned that it was Mr. Hartzog that was hurt?
Point out potential inconsistency	Your statement then is not based on anything specific that you saw?
Probe for witness bias from personal incentive	Your daughter and the plaintiff's daughter are friends, aren't they?
Repeat/rephrase question for clarity	You're sure it was my client, Mr. Roberts
Condescending counsel	You wanted praise for donating the money, right?

Table 2: Question tactic categories

[4] https://law.indiana.edu/instruction/tanford/web/reference/07cross.pdf

3. Methodology

We consider the following approaches for **CourtQG**. In Sec 3.1, we describe a vanilla approach for generating court questions based on court examination dialogue history. To go beyond basic sequence-to-sequence modeling, we discuss in Sec 3.2 approaches to acquire a deeper natural language understanding of the court questions, including agenda explanations and question type. In Sec 3.3, we then propose how to leverage agenda and question type predictions to further improve our primary task of court domain question generation. Finally, in Sec 3.4, we investigate approaches to generate court questions preemptively without dialogue context, to mimic the preparation process of court trials, based on background documents (*e.g.,* complaint files).

3.1. Dialogue-based Question Generation

Information-probing strategies may evolve based on the previous witness response, as court examination dialogue exchanges are dynamic in nature. As an intuitive first step, we explore a straightforward sequence-to-sequence approach to generate the next question (Q) that a lawyer is likely to ask in the trial, directly from court examination dialogue history. This serves as a key foundation for assistive tools that can be used in court cases.

Model We finetune a pretrained BART-based language model generator (Lewis et al., 2020) on court transcripts, which provide natural labels for the next question that a lawyer will ask based on the preceding court examination dialogue history, as depicted in Fig 3.

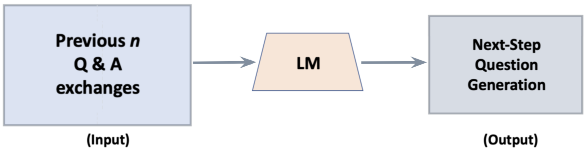


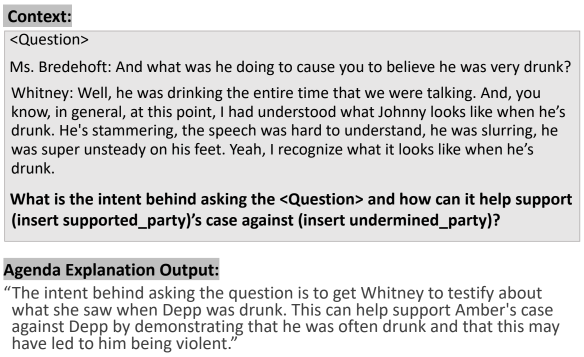
Figure 3: Vanilla baseline for dialogue-based **CourtQG**, which we extend upon in upcoming subsections.

3.2. Towards Deeper CourtQG Understanding on the Agenda and Question Types

To improve transparency and provide greater insights for end-users of **CourtQG** tools, we propose to explore court domain question understanding as auxiliary subtasks. Specifically, we consider *question type prediction*, defined as predicting the discrete category of information element likely to be asked next. We also consider *agenda explanation generation*, defined as generating the natural language explanations that describe the motivation or game-plan behind the question that will be asked next, in court examination. By modeling agenda and question type as auxiliary subtasks and incorporating the predicted information into language model input, CourtQG aims to generate questions that are more contextually relevant and transparent. Consequentially, the generated questions become more targeted and effective, aligning better with the objectives of legal counsel and the overarching goals of court examinations.

Model For question type prediction, we finetune a BART language model with a classifier head, to perform classification of the information element probed based on previous dialogue exchange context. For agenda explanation behind court examination questions, we finetune a sequence-to-sequence BART language model to generate agenda explanations.

To support such training, we derive silver-standard agenda explanation labels based on zero-shot GPT-3 prompting which has demonstrated strong capabilities for a large variety of reasoning tasks (Brown et al., 2020). We find that by feeding in any question asked in court along with the witness response for context, followed by a task description that inquires about the underlying agenda, GPT-3 can provide us reasonable agenda explanations. As we see in Fig 4, an example agenda explanation of questioning a witness about the accused party's drunken state is to tease out conditions that "may have lead to [him/her] being violent".



Context:

<Question>

Ms. Bredehoft: And what was he doing to cause you to believe he was very drunk?

Whitney: Well, he was drinking the entire time that we were talking. And, you know, in general, at this point, I had understood what Johnny looks like when he's drunk. He's stammering, the speech was hard to understand, he was slurring, he was super unsteady on his feet. Yeah, I recognize what it looks like when he's drunk.

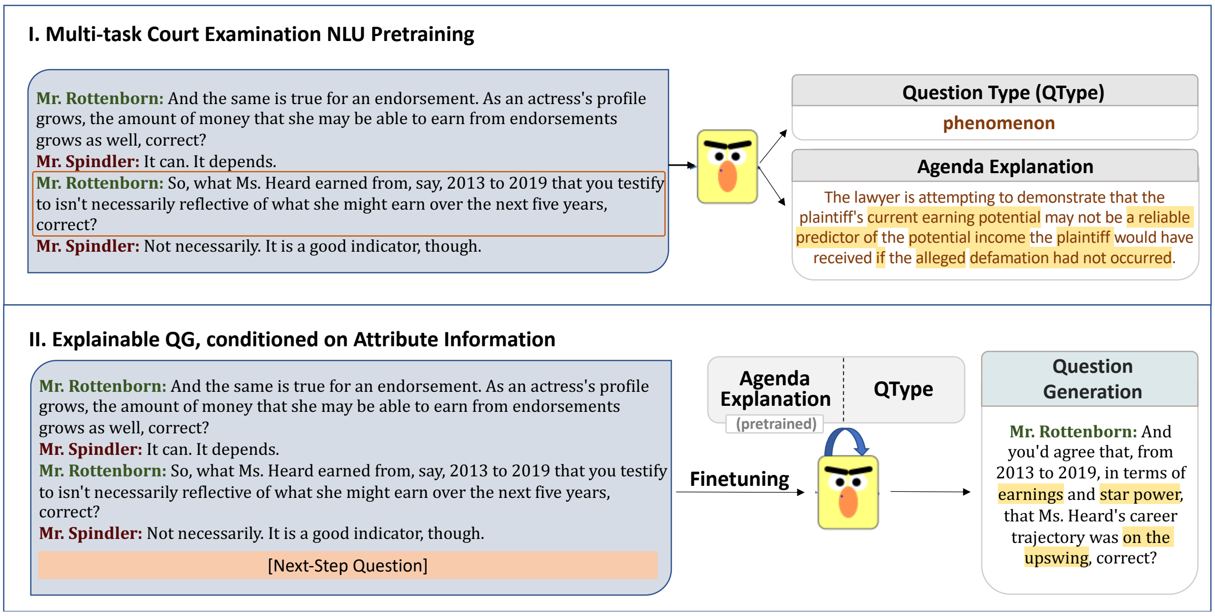
What is the intent behind asking the <Question> and how can it help support (insert supported_party)'s case against (insert undermined_party)?

Agenda Explanation Output:

"The intent behind asking the question is to get Whitney to testify about what she saw when Depp was drunk. This can help support Amber's case against Depp by demonstrating that he was often drunk and that this may have led to him being violent."

Figure 4: An illustration of the prompt template for court question agenda explanation, and "oracle" label from GPT-3 response.

It is worthwhile to note that while GPT-3 provides us good quality (silver-standard) agenda explanationss, it is only accessible through an API, which allows people limited flexibility for further experimentation and customization. The merits of our



Mr. Rottenborn: And the same is true for an endorsement. As an actress's profile grows, the amount of money that she may be able to earn from endorsements grows as well, correct?
Mr. Spindler: It can. It depends.
Mr. Rottenborn: So, what Ms. Heard earned from, say, 2013 to 2019 that you testify to isn't necessarily reflective of what she might earn over the next five years, correct?
Mr. Spindler: Not necessarily. It is a good indicator, though.

The lawyer is attempting to demonstrate that the plaintiff's current earning potential may not be a reliable predictor of the potential income the plaintiff would have received if the alleged defamation had not occurred.

Mr. Rottenborn: And you'd agree that, from 2013 to 2019, in terms of earnings and star power, that Ms. Heard's career trajectory was on the upswing, correct?

Figure 5: Our joint learning and inference framework for question generation. First, a Seq2Seq language model is pretrained on question type understanding and agenda explanation. Then, for the primary task of discourse-aware question generation, we finetune this language model to generate court examination questions, while further feeding the question type and agenda as auxilliary input. The parts in the question generation text aligned with the agenda are highlighted.

work in training local BART-based language models to perform agenda explanation generation is not limited to enhancing transparency about court questions. We also contribute language model checkpoint, that is tailored for court domain reasoning through agenda explanation generation, as an accessible resource for further adaptation, such as for the primary **CourtQG** task.

3.3. Leveraging Question Type and Agenda Understanding for CourtQG

Now, we consider two approaches for leveraging the understanding of question type and underlying agenda to court domain boost question generation. The first line of approach leverages the implicitly learned reasoning capability. In particular, we consider the following variations on top of the vanilla BART baseline:

- **BART+PT$_{QT}$** We pretrain BART first on question type prediction, and then finetune it for CourtQG.

- **BART+PT$_{AG}$** We pretrain BART first on agenda explanation generation, and then finetune it for CourtQG.

The second line of approaches leverages auxiliary data (AD), from the secondary subtasks of predicting question type and underlying agenda, for question generation. In particular, we consider the following variations on top of the vanilla BART baseline, for CourtQG training and evaluation:

- **BART+AD$_{QT}$** and **BART+AD$_{QT'}$** We condition BART with the ground truth (+AD$_{QT}$) and predicted (+AD$_{QT'}$) information type of the next time-step question, respectively.

- **BART+AD$_{AG}$** and **BART+AD$_{AG'}$** We condition BART with the oracle and predicted agenda explanations of the next time-step question, respectively.

Finally, we consider a framework that combines all the auxilliary training and information extraction components together for tackling CourtQG, as illustrated in Figure 5. We refer to our ultimate framework as **BART+{PT,AD}** for the remaining parts of the paper.

3.4. Cold-Start Question Recommendation

Finally, an important real-world scenario for court examination question generation is to prepare a set of important questions in advance even before the trial, for asking a witness. This is considered the *cold-start* scenario, in which there is no preceding

QA exchange to cue question generation. In such case, question generation relies heavily on commonsense knowledge of the social/moral/ethical rules-of-thumb (Forbes et al., 2020), and the background knowledge from court case complaint files, rather than dialogue context.

Model For this cold-start case study extension, we investigate several sequence-to-sequence language generation models that are competitive in zero-shot question generation:

- **GPT3-DAVINCI3** – For prompting, we utilize the following prompt template:

  > *Given the following court case background:* {complaint_point}
  > *And the witness description:* {witness_info}
  >
  > *List some questions to ask the witness in court examination:*

- **LLAMA2-70B** – This is an open-sourced foundation model (Touvron et al., 2023), which we feed in similar prompt template as above, for cold-cache CourtQG.
- **BLENDER-CONVAI** – This is the open-domain chatbot model trained on several conversational AI datasets (Roller et al., 2021). We utilized the 1B distilled version[5].
- **T5-SQUAD** – This is a T5-base question generation model[6] that has been pretrained on the Wikipedia-based SQUAD QA dataset (Du and Cardie, 2017; Rajpurkar et al., 2016).

We employ an automated procedure to extract the top k complaint points from the background complaint document of each court case. These documents generally adhere to a standardized format, where the main points of the complaint are listed as numbered bullets on separate lines. This structure allows for a straightforward parsing process, where we locate the substring following "\n1", "\n2", and so forth. We then feed each extracted complaint point as input to the language models above to derive relevant question generations.

4. Evaluation

To measure the quality of courtroom question generation, we utilize standard token-based metrics and semantic-based metrics, with the actual questions asked in our corpus of court cases as ground truth. The specific metrics include:

- **Traditional N-gram token matching** - ROUGE scores (Lin, 2004)
- **Semantic Similarity (SimSc)** - The court examination domain is unique in that the legal counsels come with a prepared set of questions to probe witnesses. Hence, questions generated from machine, given a dialogue context, should be rewarded credit if it matches the actual questions asked at any future time step. To actualize this, we consider the cosine similarity between embeddings of generated and ground truth question pairs, computed from a BERT-base sentence transformer model[7] trained on Quora domain question deduplication (DataCanary, 2017).
- **Entailment with Ground Truth Questions in Future Time Steps (EntSc)** - We also use the pretrained FACEBOOK/BART-LARGE-MNLI model backbone to score entailment between text pairs.
- **Agenda Score (AgSc)** - Whether the questions generated properly reflect the stance *for* or *against* a witness called by a court case party. For this metric, we rely on GPT4 automatic assessment (Liu et al., 2023), and input into the powerful language model backbone a prompt asking whether a set of generated questions better reflect "direct examination" or "cross examination", taking the token logprobs in text output as the computed score of agenda awareness.

5. Experiment and Result

We first detail the experiment and result for dialogue-based CourtQG, along with the two auxiliary tasks of question type prediction and agenda explanation, in Sec 5.1. Then, we detail the experiment and result for cold-cache CourtQG in Sec 5.2.

5.1. Dialogue-based COURTQG

We utilize a pretrained FACEBOOK/BART-LARGE as language model backbone, and train using a standard AdamW optimizer setup with 5e-5 learning rate. Each experimental run consists of 10 training epochs, and the model checkpoints with lowest validation loss are utilized for evaluation.

5.1.1. Performance Result on Auxiliary Tasks

We performed human assessment to assess the quality of questions and agenda explanations generated. We ask three human annotators to rate the

[5] https://huggingface.co/facebook/blenderbot-1B-distill

[6] https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap

[7] https://huggingface.co/deepset/quora_dedup_bert_base

	Rouge-L	SimSc	EntSc	AgSc
BART$_{\textbf{baseline}}$	20.0	31.2	41.6	64
BART+AD$_{\textbf{QT}}$	23.0	32.1	42.1	69
BART+AD$_{\textbf{QT}'}$	23.1	32.0	42.8	67
BART+PT$_{\textbf{QT}}$	23.2	32.4	41.5	65
BART+AD$_{\textbf{AG}}$	25.7	33.7	44.2	73
BART+AD$_{\textbf{AG}'}$	20.0	30.3	45.4	72
BART+PT$_{\textbf{AG}}$	19.9	28.5	42.0	70
BART+{PT,AD}	23.4	34.9	47.7	76

Table 3: Main question generation experimental results. The second box in conditioned on auxilliary information for information priming. The third box is pretraining on the auxilliary tasks of question type and agenda prediction. Note that all method variations presented in the tables are proposed by us for the novel CourtQG task setting.

Task	Method	Relevance	Convincing.
QG	BART+QType	4.5	3.4
AG	BART (Pred)	4.3	3.2
AG	GPT3 (Oracle)	4.4	3.8

Table 4: Human assessment on question generation (QG) and agenda explanation (AG).

generated questions and agenda explanations on a Likert scale of 1-5, in terms of relevance and convincingness. Table 4 shows the qualitative results from the average of ten data samples. Based on the human assessment results, we observe that generated questions or agenda are generally very relevant (4+), while the convincingness level (3-4) can still be improved.

5.1.2. Courtroom Question Generation

Finally, we compare question generation results utilizing pretraining and auxiliary data training objectives, against the vanilla sequence-to-sequence language model baseline. In Table 3, we show that pretraining on agenda explanation improves on language model performance over vanilla question generation. Conditioning the question generation on question type and agenda type leads to further improvement, such as 15% in relative ROUGE score.

5.2. Cold-Start Question Generation

Table 6 shows a qualitative comparison of cold-start question generation approaches. In general, zero-shot prompting with the large language model, GPT3, generates questions that best align with real-world court examination questions in terms of relevance, specificity, and tone/framing. The BLENDER model pretrained in the conversation domain (Roller et al., 2021) creates excessively empathetic and informal questions, inserting personal opinions such as *"Oh wow, I didn't know that. I wonder if..."* which does help elicit or concretize constructive points to the trial. Moreover, the T5 model pretrained in the SQUAD domain generates overly basic and logistically-oriented questions, such as about the date of *"opening statements"* for court cases, reflecting the news summarization nature of the Wikipedia source domain, which differs from the target domain of direct human-human interaction in court examination questions.

	SimSc	EntSc	AgSc
GPT3-prompting	0.32	0.52	0.71
LLAMA2-prompting	0.30	0.49	0.68
BLENDER-CONVAI	0.19	0.29	0.62
T5-SQUAD	0.22	0.55	0.56

Table 5: A comparison of cold-start QG results through quantitative evaluation metrics.

5.3. Remaining Challenges and Discussion

The usage of knowledge in our approaches is limited to the textual information presented in the transcribed court dialogues and complaint file. In the real-world, some evidences may be present in the form of multimedia input (*e.g.,* image, audio, etc.). Furthermore, the legal team will likely do more extensive background probing on the opposing party and witnesses, such as gathering additional evidence through private investigators.

6. Related Work

Question Generation: Previous work on question generation mostly began with a focus on generating reading comprehension style questions, i.e.,

	Domain	Example Questions Generated	Closest G.T. Question Match	Score
GPT3	LLM Prompting	So, when you returned to Los Angeles, what, if anything, took place with any relationship with Mr. Depp?	What was your relationship with Johnny Depp like after the alleged incident?	(0.51, 0.97)
LLAMA2	LLM Prompting	Opinions on the validity and reliability of the evidence presented	You don't agree that that is the gold standard assessment for reliable, accurate, psychiatric diagnosis?	(0.27, 0.92)
BLENDER	ConvQG	Wow, that's interesting. I wonder if she was a victim of domestic violence in the past?	Your father hit you and your sister at times, right?	(0.40, 0.99)
T5	SQUAD QG	What was Depp's private physician's rank?	Dr. Kipper, do you recognize this document?	(0.25, 0.13)

Table 6: Zero-shot cold-cache question generation result examples, along with their (SimSc,EntSc) scores.

questions that ask about information present in a given text (Duan et al., 2017; Gangi Reddy et al., 2022). Rao and Daumé III (2018) began to introduce the task of clarification question generation in order to ask questions about missing information in a given context. However, unlike our work, these approaches still suffer from estimating the most useful missing information.

Meanwhile, work on conversational question answering has explored the aspect of question generation or retrieval (Choi et al., 2018; Aliannejadi et al., 2019). Qi et al. (2020) especially focuses on generating information-seeking questions while Majumder et al. (2020) propose a question generation task in free-form interview-style conversations. From a broader scope, our work also draws inspirations from goal-oriented dialogue systems (Ham et al., 2020), in particular those involving theory-of-mind detection (Zhou et al., 2023), negotiation understanding (Yang et al., 2021), and counterspeech reasoning (Gupta et al., 2023), culminating these concepts for the previously unexplored courtroom examination question generation domain.

Legal Domain NLP: The legal domain provides a wide range of different tasks in which NLP techniques can and have been used (Zhong et al., 2020). Such tasks include legal document classification (Limsopatham, 2021), information extraction (Bommarito II et al., 2021), question answering on policies (Ravichander et al., 2019), court view generation (Wu et al., 2020), judicial decision-making (He et al., 2024), and case summarization (Polsley et al., 2016). The Legal General Language Understanding Evaluation (LexGLUE) benchmark (Chalkidis et al., 2022) uniformizes several legal NLP datasets oriented around the multi-label classification of law and contract documents (Chalkidis et al., 2021) and multiple-choice QA for relevant case holdings. Chalkidis et al. (2020) further explore transfer learning for the domain adaptation of pretrained language models onto legal corpora. To the best of our knowledge, our work is the first to investigate question generation for the legal domain.

7. Conclusion and Future Work

In this paper, we explored a novel and interesting problem of court examination question generation, which we also refer to as CourtQG. We formalized the problem domain through three relevant tasks: dialogue-aware QG, cold-start QG, and question type understanding (modeling the type of information probed and generation of explanation on underlying agenda). In addition, we benchmarked the performance of relevant baselines, and discussed insightful observations. CourtQG represents a larger class of agenda-driven and interaction-aware problem space for question generation and information probing. In future work, we plan to explore the transferability of methodology for other agenda-driven, discourse-aware question generation tasks such as for interviews, etc. We also aim to study human-in-the-loop setting for iteratively improving agenda-driven question generation to align better with human intents, such as incorporating a more comprehesnvie set of *latent persona* reasoning (Sun et al., 2023) for courtroom domain agenda-awareness.

8. Ethics Statement & Broader Impact

The development and implementation of AI-assisted question generation systems in the court domain introduce various ethical considerations. These considerations revolve around ensuring fairness, transparency, privacy, and accountability in the use of such technology. It is essential to address these ethical concerns to promote responsible and ethical AI practices within the legal system.

In particular, key ethical considerations include:

- **Fairness and Bias**: AI models for question generation must prioritize fairness and mitigate bias by addressing potential biases in training data and conducting regular audits and testing to identify and rectify any biases, including historical disparities in legal outcomes.

- **Accuracy and Reliability**: In the court domain, ensuring fact-grounded reliable AI-generated questions is vital, requiring extensive testing, validation procedures, and regular updates to enhance the precision and trustworthiness of the questions generated by the AI systems.

- **Privacy and Confidentiality**: Robust privacy and security protocols are crucial in AI-assisted question generation for the court domain, safeguarding sensitive information, court documents, case details, and personal data through access controls, data encryption, secure storage, and strict adherence to data protection regulations and guidelines.

- **Transparency and Explainability**: Transparency and explainability are vital for AI-assisted question generation systems, ensuring trust and accountability. Users and stakeholders should understand the system's operations, including data sources, algorithms, and decision-making, while clear explanations and transparency about limitations facilitate meaningful human oversight and intervention.

- **Human-Centered Approach**: AI-assisted question generation systems should be human-centered, complementing legal professionals rather than replacing them, with a focus on prioritizing human involvement, expertise, and collaboration between AI developers, legal experts, and stakeholders to ensure alignment with legal requirements and ethical standards.

The ethical considerations outlined above provide a framework for responsible development and deployment of AI-assisted question generation systems in the court domain. By addressing fairness, accuracy, privacy, transparency, and maintaining a human-centered approach, we can promote the ethical use of AI technology while upholding the principles of justice and the rule of law. It is essential to engage in ongoing discussions, collaborations, and regulatory efforts to ensure that AI systems in the court domain serve the interests of justice while maintaining the highest ethical standards.

Bibliographical References

2023. United states district court northern district of california transcripts / court reporters. Accessed on January 1, 2023.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on Big Data Law*, pages 216–227. Edward Elgar Publishing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Lili Jiang Meg Risdal Nikhil Dandekar tomtung DataCanary, hilfialkaff. 2017. Quora question pairs.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Yi R. Fung, Kevin Small, and Heng Ji. 2022. A zero-shot claim detection framework using question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6927–6933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Simucourt: Building judicial decision-making agents with real-world judgement documents.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Nut Limsopatham. 2021. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 210–216, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Revanth Gangi Reddy, Yi R. Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, and Heng Ji. 2023. Smartbook: Ai assisted situation report generation. In *arXiv*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kevin Ros, Maxwell Jong, Chak Ho Chan, and ChengXiang Zhai. 2022. Generation of student questions for inquiry-based learning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 186–195.

Chenkai Sun, Jinning Li, Yi R. Fung, Hou Pong Chan, Tarek Abdelzaher, ChengXiang Zhai, and Heng Ji. 2023. Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 681–693, Online. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, Toronto, Canada. Association for Computational Linguistics.