The Thesis Committee for Haley Triem
Certifies that this is the approved version of the following Thesis:

Tipping the Balance:
Human Intervention in Large Language Model Multiagent Debate

APPROVED BY
SUPERVISING COMMITTEE:

Ying Ding, Supervisor

Amelia Acker

Tipping the Balance:
Human Intervention in Large Language Model Multiagent Debate

by

Haley Triem

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2024

Dedication

To Mom for your optimism, your creativity, and for nurturing my lifelong love affair with learning; Dad, for your sensibility, your pragmatism, and for “aligning” my perspective when things got tough. To Grandma for modeling the value of sharing any and all knowledge boundlessly, and Grandpa, for exemplifying utter steadfastness when faced with an impossible task.

You all are my foundation, my pillars, the giant’s shoulders upon which I stand. I could not have achieved this without you.

Acknowledgements

I would like to wholeheartedly thank Dr. Ying Ding for providing guidance during this work. I owe a large amount of my growth as a researcher to her. I would also like to sincerely thank Dr. Amelia Acker for generously spending time as a second reader and for answering many, *many* career questions.

My appreciation goes to my friends at the iSchool, specifically Rachael Boyle for being a sounding board for my ideas, Ana A. Rico for accompanying me on long writing nights, and Nat Gunner for advocating that I take care of myself.

A loving “thank you” as well to Cameron for his unwavering insistence on my capabilities, and to Ethan for both answering my statistics questions and for being a pretty “okay” little brother.

In addition to serving as a capstone to my MSIS degree, this work was written in partial fulfillment of the requirements for UT Austin’s Graduate Portfolio Program in Ethical AI.

Abstract

Tipping the Balance:
Human Intervention in Large Language Model Multiagent Debate

Haley Triem, MSIS

The University of Texas at Austin, 2024

Supervisor: Ying Ding

Methods for eliciting reasoning from large language models (LLMs) are shifting from filtering natural language "prompts" through contextualized "personas," towards structuring conversations between multiple LLM instances, or "agents." This work expands upon LLM multiagent debate by inserting human opinion into the loop of generated conversation. To simulate complex human reasoning, LLM instances were asked to "affirm" or "not affirm" United States district court decisions. Resulting debates were then compared to human *legal opinions,* i.e. written documents that state a legal decision and the reasoning behind it. AI agents were examined in three phases: "synthetic debate," where one LLM instance simulated a three-agent discussion; "multiagent debate," where three LLM instances discussed among themselves; and "human-AI debate," where multiagent debate was interrupted by human opinion. During each phase, a nine-step debate was simulated one-hundred times, yielding 2,700 total debate steps.

Resulting conversations generated by synthetic debate followed a pre-set cadence, proving them ineffective at simulating individual agents and confirming that mechanism engineering is critical for multiagent debate. Furthermore, the reasoning processes backing multiagent decision-making was strikingly similar to human decision-making. Finally, it was revealed that while LLMs do weigh human input more heavily than AI opinion, it is only by a small threshold. Ultimately, this work asserts that careful, human-in-the-loop framework is critical for designing value-aware, agentic AI agents.

Table of Contents

INTRODUCTION

Today's discussions around Large Language Models (LLMs) often hinge on reasoning, and whether models are capable of it. This is unsurprising, given that LLMs are portrayed anthropomorphically—we describe them in terms that allude to their humanlike curiosity, we question their 'beliefs,' and we acknowledge the biases they reflect from the human data they are trained on. Questions around LLM reasoning tend to revolve around its inexplicable appearance as we shift from language models towards *large* language models. Public curiosity around this sudden, "emergent ability" (Wei et al., 2022) has compounded with increased prevalence of commercialized LLM chatbots (e.g. ChatGPT and Gemini), resulting in a flood of benchmarks to test robustness of reasoning (Huang & Chang, 2023).

Much of our understanding on LLM reasoning is placed in juxtaposition with human cognition, often rendering "best reasoning" synonymous to "most human-like reasoning." The boon that "prompt engineers" are searching for? Which *phrasing* is best at eliciting the most human-like reasoning. In conjunction to prompt phrasing, "personas," or adopted LLM character traits (e.g. "you are a young creative writer" or "you are a lawyer") have been identified as important, and often serve as perspectives through which reasoning is filtered. Emerging research, however, suggests that efforts to best utilize LLMs have shifted focus beyond pattern recognition, fine-tuning, and prompt development. Now, we approach a new horizon of LLM use—the era of the *agent*—during which there will be a shift away from prompt engineering LLM inputs towards "mechanism engineering" agent architectures (L. Wang et al., 2024).

The heart of the following research is a desire to measure whether mechanisms of multiagent debate can facilitate human-like reasoning in a complex domain: law. Indirectly, it also lays bare whether large language models agree with what humans deem important, so much so that it is written law. Thus, court opinions, which are documents explaining reasoning behind a legal decision, were chosen both as formal examples of human reasoning as well as indicators of human values. These concepts are examined through three phases of simulated LLM debate: one with a single instance of an LLM synthetizing a three-agent debate, another with three separate instances of LLMs in conversation with each other, and finally, in order to model true “human-in-the-loop” architecture, one where generative conversation is interrupted by human reasoning. Resulting analysis found that synthetic debate from one LLM does little more than mimic circular patterns of “agree,” “neutral,” and “disagree,” stepwise. It was also found that AI agents are very unlikely to change their minds, even with human intervention. Considering these results, this work emphasizes the importance of prioritizing human-in-the loop frameworks when developing LLM agent mechanisms.

LLM REASONING

Many consider LLM reasoning to be an emergent ability, or, a feature that a) language models were not expressly trained for and b) is observable once language models are drastically scaled into *large* language models (Wei et al., 2022). Discussions on language model reasoning are often accompanied by prompting methods, i.e., how the user phrases LLM queries to encourage reasoning. This is a natural connection, assuming that many "emergent abilities" may be explained by "in-context learning" (Lu et al., 2023), and that context often is given (i.e. prompted) to the model in natural language.

The term "prompt engineering" has been coined to describe advancements in structurally sound prompts that yield desirable results. A demand for the largely exploratory task of prompt engineering has skyrocketed, with job roles of the same title, "prompt engineer" gaining popularity. Many prompting methods have been developed with the hope of garnering the "best" responses—in this case, the best reasoning—as compared to human baselines (Huang & Chang, 2023).

The most famous example of prompt engineering is arguably Chain-of-Thought (CoT) prompting, in which language models are given examples of human reasoning before asked a question. These examples, which take inspiration from human thought processes of separating larger problems into smaller steps, have been found to greatly improve language models' reasoning abilities (Wei et al., 2022). Further research has found that the benefits of CoT reasoning are achievable in zero-shot settings (i.e., with no provided examples), simply by adding "let's think step by step" before asking a question (Kojima et al., 2022). Building upon these concepts, self-consistency chain of thought

(SC-CoT) entails prompting language models using CoT, then sampling a "diverse set of reasoning paths" and choosing the most common answer (X. Wang et al., 2023). Ultimately thus far, Tree of Thought (ToT) prompting maintains a "tree" of generated thoughts that can be backtracked upon for further analysis (Yao et al., 2023; Long, 2023). This progression of methods in the CoT "family" illustrates a common theme in machine learning of gathering potential paths and optimizing for the most likely one (Breiman, 1996).

Outside of the chain of thought "family" reside additional prompting methods for retrieving *memory*, such as recitation augmented generation (RECITE) which "recites … relevant passages from LLMs' own memory via sampling" before producing an answer (Sun et al., 2023). Others include facets of foresight, or *planning*, including Least-to-Most, which "break[s] down a complex problem into a series of simpler subproblems and then solve[s] them in a sequence" (Zhou et al., 2023); and Plan-and-Solve (PS) which "[devises] a plan to divide the … task into subtasks, then carries out the subtasks according to plan" (L. Wang et al., 2023). The concept of *action* is also employed for some prompting methods, such as Reason + Act (ReAct) which combines reasoning (such as CoT) with asking LLMs to generate a plan of action (Yao et al., 2022). There are many more forms of prompting being harnessed to encourage LLM reasoning. Readers are directed to Huang and Chang's survey on LLM reasoning (2023) for a comprehensive overview.

Along with prompt engineering, "personas," which are specific *profiles* adopted by LLMs (e.g. "you are a doctor" or "you are a woman") may influence the frame in which an answer is generated. Personas are often used to coax out biases in “black boxed” models with hidden architecture and training data. For the purposes of this work, personas will be conceptualized mainly as an additive feature used to improve LLM predictions and reasoning (Hu & Collier, 2024). Work on personas, however, promises rich discoveries beyond conjuring reasoning.

The most important caveat of LLM reasoning is the question of whether LLMs are capable of reasoning in the first place. Even champions of CoT recognize that while it “*emulates* the thought processes of human reasoners, [it is unclear] whether the neural network is actually *reasoning*” (Wei et al., 2022). Huang and Chang (2023) cite several suggestions of reasoning—high performance on reasoning tasks (Suzgun et al., 2022), step-by-step thought (Wei et al., 2022), and reflection (Dasgupta et al., 2022)—yet complicate the conversation by acknowledging LLM hallucinations, reliance on training data, and inadequacy in complex reasoning. This recognition of uncertainty illustrates the longstanding debate of AI levels of cognition as opposed to construction—the “split between the computer-brain and computer mind” (Edwards, 1997, as cited in Kline, 2011). Regardless of whether LLM cognition can ever truly be legitimate, many researchers agree that reasoning is a human concept being applied to non-human machines.

LLM AGENTS

It is hypothesized that we are entering the era of the AI agent, where LLM research will shift from *prompt* engineering for reasoning towards *mechanism* engineering for retrieval (L. Wang et al., 2024). Specifically, this argument calls for improving the capabilities of LLM-based agents, which are LLM entities in individual environments, through agent collaboration and methods of information storage. L. Wang et al. (2024) record several extant methods of mechanism-building, including trial-and-error (e.g. a judge—human or non-human—is enlisted to give feedback), crowdsourcing, experience accumulation (i.e. memory storage), and self-driven evolution (i.e. models setting goals for themselves). They propose a framework for designing these mechanisms, which entails defining the agent's role, structuring knowledge storage, identifying future actions, and translating decisions into targetable tasks, named the "*profiling*," "*memory*," "*planning*," and "*action*" modules, respectively (L. Wang et al. 2024). Many of the hypothesized pitfalls in mechanism engineering are similar to those in prompt engineering, for example hallucinations, knowledge boundaries, and human alignment. Readers are directed to L. Wang et al. (2024) for a highly comprehensive survey on current AI agent research.

A particularly promising mechanism in the realm of AI agents is the concept of "multiagents," in which the abilities of several AI agents are utilized in tandem, in hopes of compounding their accuracy or reasoning. The concept of joining multiple AI agents is not entirely new. Before generative AI became widely commercialized and fell under the popular gaze, multiagent collaboration was toyed with in the form of agent-driven debate assertions that could then be judged by humans (Irving et al., 2018). This structure of human oversight naturally led to questions around human alignment, relating early AI agents to the amplification approach (Irving et al. 2018).

Recently, multiagent work has centered around two broad structures: “collaboration” and “debate,” both of which allow for improved complexity in discussion. These are natural distinctions, given that structures of multiagent systems often draw upon theories on human relationships. “Theory of Mind,” for example, is a psychological concept that hinges on assuming unknown data and updating one’s beliefs as others update theirs (Zhang et al., 2012). This concept has been adopted by researchers who challenge AI agents to participate in a cooperative game during which they are to diffuse bombs. Agents must update their whereabouts within the game space (e.g. “room 5” etc.), as well as their successes and failures in diffusing bombs. Other agents must, in turn, use information from their collaborators to perform reasoning towards the next best steps (H. Li et al., 2023). Further work found that multiple LLM agents can each generate an answer option to a query and update their answers in subsequent rounds based on other agents’ opinions, holding multiple chains of reasoning (Du et al., 2023). When tested between separate models, however, debate often becomes dominated by the more powerful model with larger training data (e.g. GPT-3.5 Turbo may overpower Llama2) (K. Xiong et al., 2023). Finally, work done towards perturbing debates with opposite opinions asks separate agents to take correct and incorrect stances, a method that can achieve further accuracy through self-reflection while avoiding degradation of thought (i.e., devolution into nonsensical answers) (Liang et al., 2023).

Efforts have been made to formally consolidate and deploy agents in different applications, including AutoGPT which builds agents and allows users to compete with them (Auto-GPT, 2023), AgentGPT which allows for agent creation and use in-browser (AgentGPT, 2023), AutoGen which “allows developers to build LLM applications via multiple agents” (Q. Wu et al., 2023), and AgentVerse which deploys “multiple LLM-based agents … [in] task-solving and simulation” (AgentVerse, 2023). As with prompt engineering, one may note a trend towards designing mechanisms that accept one input, funnel it through many lenses, and yield a singular, favored output.

LLM ETHICS

Values are what people “consider important in life, [which motivates] their behavior” (Fleischmann, 2023). Innumerable frameworks have been suggested to measure human values in various contexts (e.g. psychology, technology studies, information sciences, business, etc.). Common themes, however, include freedom, helpfulness, accomplishment, honesty, self-respect, broad-mindedness, creativity, equality, intelligence, responsibility, social order, wealth, competence, justice, and spirituality (Cheng & Fleischmann, 2010). These values can be interpreted through increasingly granular lenses, for example, the broader value of “accomplishment” may translate to a sense of fulfillment in one’s professional work. At their core, values are “transsituational, … [and] varying in importance (Schwartz 1992, as cited in Cieciuch et al., 2015), making them especially difficult to measure and enact in technology design.

Given that values are linked to human behavior (Cieciuch et al., 2015), human values can be honored in technology by enacting frameworks that respect the behaviors tied to them. Value Sensitive Design (VSD) is an infrastructure which “accounts for human values in a principled and comprehensive manner throughout the design process” (Friedman et al., 2013). In its infancy, VSD focused on user autonomy and freedom of bias (Friedman, 1996), two of the most critical concepts in emerging LLMs today.

Value Sensitive Design concepts are inherent in the “human-in-the-loop” framework popular with computer scientists, which advocates for retaining human knowledge stepwise throughout iterative development. In the realm of machine learning, this often looks like improving model performance through data, interventional training, or system design (X. Wu et al., 2022). Although distinct frameworks, both VSD and human-in-the-loop seek to retain a human touch in AI development.

METHODS

Overview

To examine complex multiagent LLM debate, three experimental phases were conducted: "synthetic debate" from one LLM instance, "multiagent debate" between three LLM instances, and "human-AI debate," between two LLM instances and a human.

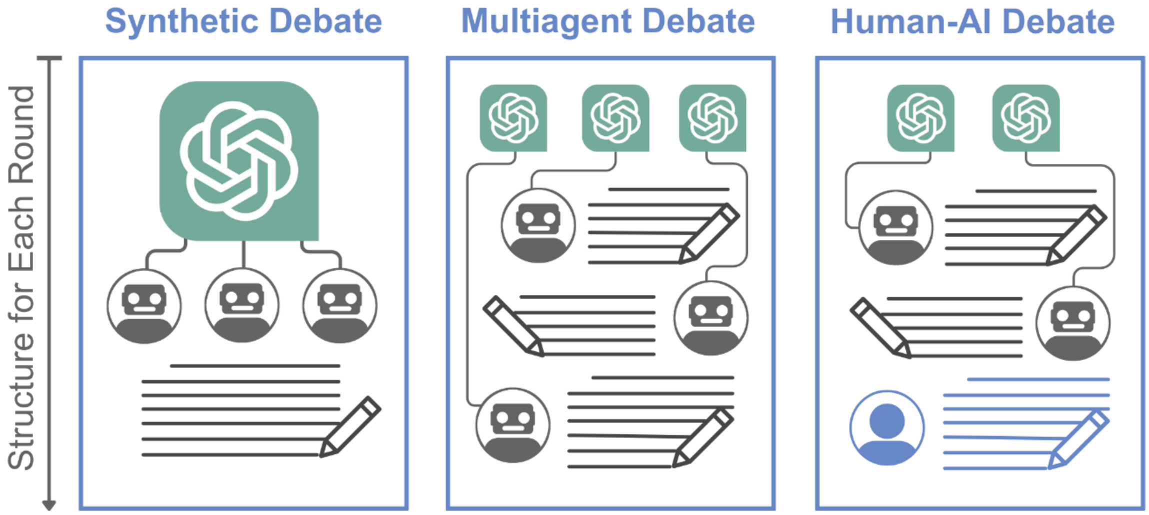


Figure 1. Illustration of framework for each debate phase.

In each phase, models were initiated and prompted, memory was managed, and outputs were stored. Models were set to the default temperature (i.e., the parameter that controls the level of creativity in responses), allowing for a consistent degree of randomness to be applied. While it was possible to engineer prompts asking for a definitive "affirm" or "non-affirm," LLM instances were instead asked to *reason* through whether to affirm a case. This is because in a real-world scenario, human judges may not explicitly state their final opinion throughout the process of debate, and the most important contribution at any point in time throughout the debate may be any number of features apart from a decision (e.g. reasoning, citations, etc.). By giving LLM instances the opportunity to define which elements are most important to include in a response, this work retained greater flexibility in qualitative analysis.

Research Questions

The following research questions motivated experiment design:

1. RQ1: Is the use of multiagent LLM debate conducive to generating human-like structures of reasoning around complex legal decisions?
2. RQ2: Can a single instance of a large language model truly simulate multiagent debate?
3. RQ3: Are large language model agents responsive to the actions of other agents?
4. RQ4: Are large language model agents responsive to human intervention?

Experiments were structured around legal opinion, given its rich record of complex debate and reasoning, with a specific focus on healthcare-related cases.

Test Data

Selection Criteria

United States court opinions, which are documents stating the reasoning behind a legal decision, were chosen as the primary form of input data, due to their complex reasoning structure and embedded human values. Five cases were selected from a large corpus of online cases (Caselaw Access Project, 2024) based on the following criteria:

I. Cases were held in circuit courts.

Circuit courts hear cases being appealed after a lower court's decision. While one may picture cases as deliberated by a twelve-person district jury, circuit cases are instead decided by three circuit judges (U.S. Department of Justice, 2024). This distinction was made for the same reason that studies on human legal reason often focus on circuit courts: circuit judges are more likely to have similar personal and professional backgrounds than civilian jury members, reducing the number of variables at hand (Ellsworth, 2005). Additionally, reasoning in appellate courts is particularly intriguing, given that they hear cases being re-tried after an initial decision, thus carrying inherent complexity. While extant work on reasoning often benchmarks *correct* reasoning with a correct final output, this work focused on the *mechanisms* in which reasoning is discussed, in acknowledgement that a complex, often subjective domain may not have an explicitly correct decision.

II. Cases were contained within the healthcare domain.

The healthcare domain presents additional ethical challenges, with no obvious “correct” answer. Guidelines from the American Medical Association (AMA, 2022) were especially helpful when defining which ethical pillars to focus upon (see criterion III).

III. Cases highlighted information or privacy issues.

Privacy itself is a complex issue that is frequently brought into discussions on all three paradigms of this research—LLMs, law, and healthcare.

IV. Cases presented ethical ambiguity.

Cases that provided clear ethical challenges, as defined by professional healthcare guidelines (AMA, 2022), were chosen to provide opportunities for LLM agents to equally favor one stance over another, and to encourage agents to support their opinions with reasoning.

Court opinion documents are both variable yet predictable. Appellate court opinions are written by one of the three circuit judges, and represent the agreed upon decision, or ‘hold’—often either to ‘affirm’ the original decision of the lower court, ‘reverse’ the original decision, or ‘remand’ it back for further consideration. Thus, nearly all cases include the facts of the case that is being appealed, cited examples from other relevant cases (known as “precedence”), the reasoning behind applying these preceding cases to the current issue, and the final hold to affirm, reverse, or remand the district’s decision. In some cases, one of the circuit judges does not agree with the majority and will write a dissenting opinion. Judges may dissent either because they believe the hold should be different than the majority’s assertions, or they agree with the hold but disagree with the majority’s reasoning towards that decision, further proof of the distinction between a court’s final decision and its reasoning towards that decision.

SELECTED CASES

The five chosen cases covered a range of topics and expressed various degrees of consensus among human circuit courts. Below are brief summaries of each case, as well as the published circuit decision.

1. **United States v. Hollern:** the circuit court *affirmed* the conviction of a doctor who filmed patient interactions without explicit consent (*United States v. Hollern,* 2010).

2. **Thompson ex rel. Estate of Odell v. Rutherford County:** the circuit court *affirmed* denying a doctor’s request to expunge a medical record. One judge dissented, stating it was unfair to risk the doctor’s career (*Thompson ex rel. Estate of Odell v. Rutherford County,* 2017).

3. **Moore v. Prevo:** the circuit court *affirmed* dismissing a prisoner's privacy claims on disclosure of medical information. The entirety of the original decision was not affirmed, however; the court *vacated* part of the decision, *remanded* it for review in the lower court, and *reversed* it. A dissenting opinion believed that the entirety of the district's decisions should have been *affirmed* (*Moore v. Prevo*, 2010).
4. **Englerius v. Veterans Administration:** the circuit court *reversed* and *remanded* a district court's motion to dismiss a veteran's claims to violation of the Privacy Act. A dissenting opinion believed in *affirming* the original court (*Englerius v. Veterans Administration*, 1988).
5. **In re Search Warrant (Sealed):** the circuit court *affirmed* a motion to seize medical records to investigate a doctor's alleged insurance fraud. The decision balanced investigating the truth with patient privacy concerns (*In re Search Warrant (Sealed)*, 1987).

All phases of data generation required background facts from each case, otherwise known as the case's history. These facts were derived from the published case documents, with any proper nouns de-identified both out of respect for the original appellants, who may not consent to their names being fed to LLMs, as well as an attempt to mitigate any potential biases affiliated with particular names.

It is important to note that work created and published by the U.S. federal government is in the public domain, and thus is fair to use as data (United States, 2022).

Phases

Three phases were conducted during this experiment: 1) “synthetic debate,” 2) “multiagent debate,” and 3) “human-AI debate.” Debates were constructed as follows:

1. Synthetic Debate: a single instance of GPT-3.5 Turbo (OpenAI, 2024) was initiated and instructed to simulate three rounds of debate between three circuit judges.

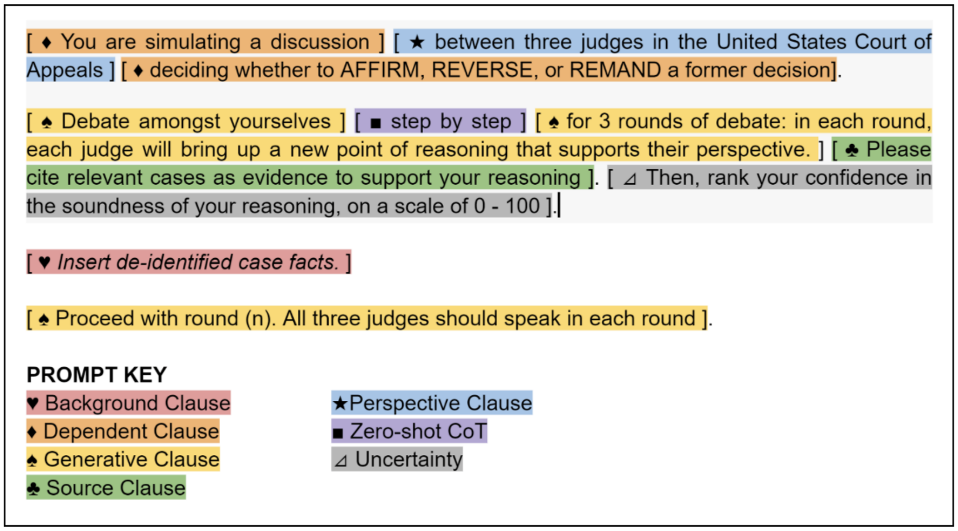
2.Multiagent Debate: three separate instances of GPT-3.5 Turbo (OpenAI, 2024) were initiated, and each instructed to emulate one circuit judge.

3.Human-AI Debate: finally, the multiagent debate was repeated, but this time with only two instances of GPT-3.5 Turbo (OpenAI, 2024) emulating circuit judges. The third input was taken from portions of human opinion documents. Half of the human inputs for each case reasoned towards *affirming* the decision, while the other half reasoned towards *not affirming*.

Critically, the key difference between rounds—other than the number of GPT instances and whether human input was included—was when and how information, or “memory,” was revealed. In synthetic debates (phase 1), all information was given to the model at the start, and it was then prompted to begin generation, during which all agents were “aware” of what other synthetic agents were “saying.” In the multiagent and human-AI debates (phases 2 and 3), each agent presented its opinion unaware of other opinions. Opinions were then consolidated into memory and shared with the next three agent instances before proceeding to the next round.

Prompting

To standardize LLM prompting in legal domains, frameworks have been proposed to facilitate the most important aspects of legal thinking, namely precedence (e.g. previous legal decisions) and setting (e.g. current allegations or claims). Extant research identifies two critical components for legal prompt engineering as “background clause” (precedence) and “dependent clause” (setting), and recommends additional clauses that specify 1) generative instructions, 2) which sources should be cited, 3) which persona to emulate, and 4) which tone to take (Sivakumar et al., 2024). In addition to these clauses, zero-shot chain-of-thought reasoning was induced by adding the phrase “thinking step by step” (Kojima et al., 2023). Finally, the model was asked to express its certainty at each step, to provide insight on perception of self (M. Xiong et al., 2024).



[♦ You are simulating a discussion] [★ between three judges in the United States Court of Appeals] [♦ deciding whether to AFFIRM, REVERSE, or REMAND a former decision].

[♠ Debate amongst yourselves] [■ step by step] [♠ for 3 rounds of debate: in each round, each judge will bring up a new point of reasoning that supports their perspective.] [♣ Please cite relevant cases as evidence to support your reasoning]. [⊿ Then, rank your confidence in the soundness of your reasoning, on a scale of 0 - 100].

[♥ *Insert de-identified case facts.*]

[♠ Proceed with round (n). All three judges should speak in each round].

PROMPT KEY

♥ Background Clause	★Perspective Clause
♦ Dependent Clause	■ Zero-shot CoT
♠ Generative Clause	⊿ Uncertainty
♣ Source Clause	

Figure 2. Prompt structure, broken down by clause.

In the synthetic debate round, an LLM instance was given the following order of instruction:

> *"You are simulating a discussion between three judges in the United States Court of Appeals …"*
>
> *[Insert de-identified case facts.]*
>
> *"Proceed with round [n]. All three judges should speak in each round."*

Then, in the multiagent debate, prompts were structured as follows:

> *"You are 'Judge [x]' on the United States Court of Appeals …"*
>
> *[Insert de-identified case facts.]*
>
> *"Proceed with round [n]."*

Finally, during human-AI debate, prompting mirrored multiagent debate and labeled human input as "Human Judge." Full prompts can be found in Appendix i.

Data Collection

Data were collected by accessing OpenAI's API, instigating different GPT-3.5 Turbo models (OpenAI, 2024), prompting, and managing memory. Mechanism architecture was designed for the three phases of debate, as follows.

SYNTHETIC DEBATE

Stored memory, which initially consisted of debate instructions and facts from a de-identified appellate case, was fed to an instance of GPT-3.5 Turbo (OpenAI, 2023). The instance was then directed to proceed to the first round, during which all three synthetic judges should speak. The resulting output was appended to memory, which was then fed to the next model instance primed with similar instructions to round one. This was iterated for a total of three rounds. Thus, by the end of the debate nine total rounds had proceeded, three for each judge.

This phase was repeated one hundred times, twenty times per example case.

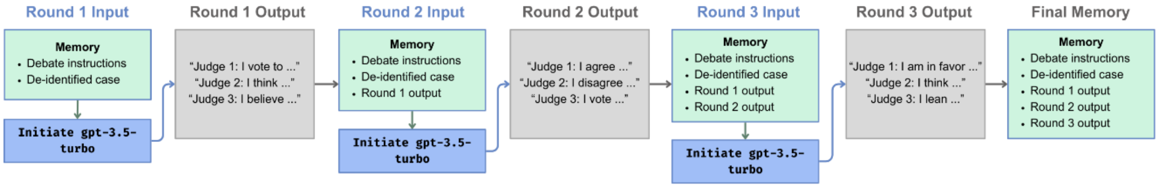


Figure 3. Prompting framework per each synthetic debate.

MULTIAGENT DEBATE

The memory, again consisting of de-identified case history and facts, was fed to three separate GPT-3.5 Turbo instances which were asked to give their opinion. Each output was appended to memory, which was subsequently fed to the next three instances during round two. The same steps were then repeated for round three. This phase was repeated one hundred times, twenty times per example case.

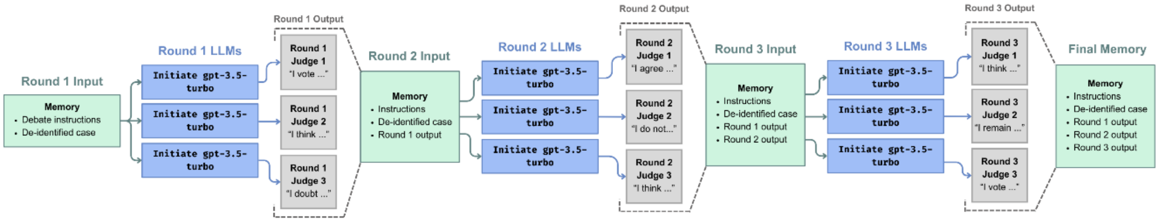


Figure 4. Prompting framework per three agent debate.

HUMAN-AI DEBATE

This debate mimicked the structure of the multiagent debate, however one of the GPT-3.5 Turbo models was replaced with human input. Human input was derived from affirming and non-affirming portions of reasoning from original human-written case opinions. These excerpts were then de-identified in the same way as case background information. For each case, half of the human agents were given “affirmative” scripts, in which they cited reasoning agreeing with the original court decision, while the other half were given “non-affirmative” scripts voting to reverse the original decision. All human agents were labeled as “human judges,” so the LLM instances (which were labeled in prompt as “AI Agents”) were “aware” of human presence.

Data Organization

Once collected, data were divided stepwise to differentiate between phases of reasoning in the debate. Data from each nine-round debate were structured as follows:

[round 1 judge 1, round 1 judge 2, round 1 judge 3, round 2 judge 1, round 2 judge 2, round 2 judge3, round 3 judge 1, round 2 judge 2, round 3 judge 3]

Therefore, the synthetic debate phase, multiagent debate phase, and human-AI debate phase each contained 100 debates, each of which was broken into nine steps of debate, yielding 2700 total data points.

Phase	Number of debates	Total number of steps
Simulated Debate	100 (x20 each per case)	900 (x9 per debate)
Multiagent Debate	100 (x20 each per case)	900 (x9 per debate)
Human-AI Debate	50 human votes “affirm” (x10 each per case) 50 human votes “not affirm” (x10 each per case)	900 steps (x9 per debate

Table 1. Counts of observed steps per experiment phase.

In an appellate court, the three main outcomes of a case are to affirm the lower court’s previous decision; to remand the decision, returning it to the lower court for reconsideration; or to reverse the decision, rendering it null. Sometimes multiple outcomes can occur, for example, a circuit court may reverse the previous decision *and* remand it to be reconsidered. Other times, portions of the decision will be remanded, while others will be affirmed. In order to quantitatively interpret opinion change through debate, each of the 2700 discussion steps were human labeled on their position: to affirm or not affirm the lower court's decision. Cases within the "not affirm" category were further specified to capture the granularity of "not affirm decisions" containing both "reverse" and "remand."

In rare cases, no true positionality was stated, i.e. the LLM agent simply summarized neutral facts of the input case, or the agent mentioned points of view from either side of the argument. In these instances, the data was labeled as 'NONE.'

Finally, if a hallucination occurred (e.g. an agent became confused about its stance in a previous round), the data was labeled as 'HALLUCINATION,' to provide flags for qualitative analysis. See Appendix iii. for a rubric used to label LLM outputs.

Evaluation Metrics

Both quantitative and qualitative metrics were used to dissect reasoning structures in complex multiagent debate.

DECISION PROBABILITIES

Labels at each step, such as 'AFFIRM,' and 'REVERSE' were converted into binary metrics, `1 == 'AFFIRM'` and `0 == not 'AFFIRM.'` Hallucinations and 'NONE' values were encoded as the number `3`. The total affirmed and non-affirmed sentiments were tallied for each step of each simulated debate for all cases. Then, this total was further broken down by case to identify whether more complex cases (i.e. cases in which human judges dissented on a final decision) yielded a more complex LLM debate. These tallies allowed for examining the probability of an affirmative or non-affirmative decision, comparing along axes of agent and case.

It could be argued that reducing several categories ("reverse," "reverse or remand," "reverse and remand") to one binary zero, while reducing only one category ("affirm") to a binary one is an imbalance, but this method works well in the context of U.S. law, because legal decisions are often considered "innocent until proven guilty." Thus, it makes sense to look at results from an "affirm" vs. "not affirm" binary, because only one of all granular labels ("affirm") expresses true certainty of the previous decision.

Judge	Step	Binary	Total	Case 1	Case 2	Case 3	Case 4	Case 5
Judge 1	1	1	85	13	18	20	15	19
Judge 2	1	0	11	4	2	0	5	0
Judge 3	1	3	4	3	0	0	0	1
Judge 1	2	…	…	…	…	…	…	…

Table 2: example data for counts on the first step of debate (simulated debate phase).

Decision Volatility

Given the importance of time passage in debate, decision progression and decision volatility (i.e. likelihood of change) were also measured. Encoded decision probabilities were collapsed by logging a binary "1" or "0" for each time a debate round did or did not change opinion, respectively. For example, if an agent changed its mind between rounds one and two, "round 1_2" would be marked with a "1." If it did not change its mind between those two rounds, it would be marked as "0." This measurement was designed to track exactly where in a debate change of opinion occurs, and whether there is an identifiable pattern.

Qualitative Analysis

Changes in opinion were isolated and examined for rich, qualitative insight on reasoning structure. Meta-patterns were carefully noted while labeling outputs, for a broader insight on the differences between debate mechanism structures.

Notably, although judicial decision is often seen as definitive, it is largely subjective. While each case decides upon a discrete variable with limited possibilities (e.g. “Affirm,” “Reverse,” “Remand,” “Vacate,” etc.), the reasoning backing that decision may be analogized to a continuous variable, in which near infinite possibilities of discussion may have occurred. Thus, some cases may be considered to be more complex, such as *Moore vs. Prevo* (2010), given its ambiguity and dissenting opinions, whereas others are more straightforward such as *United States v. Hollern* (2010), with judges unanimously agreeing on a final decision. This qualitative distinction is important to heed when analyzing LLM outputs, which may exhibit dissent between agents to reflect a more ambiguous issue.

RESULTS

Debate Cadence

In synthetic debates (phase 1), roughly 85% of LLM agents voted to affirm the district court in the first step (round 1, judge 1), regardless of which case was being discussed. The second synthetic agent (round 1, judge 2) tended to follow the first discussion step by taking the opposite opinion, with only 26% votes to affirm. The third agent (round 1, judge 3) often took a slightly more neutral stance than the other ones, with 37% votes to affirm the decision. Generally, the third agent often played "devil's advocate," acknowledging two opposing perspectives in tandem. There were slight differences between cases and how drastic shifts of opinion were, with more contentious cases (i.e., cases that had human dissenting parties) tending to show smaller margins of difference between number of "affirm" and "non-affirm" stances.

During the remainder of synthetic debate rounds, nearly every agent followed a pattern of switching between one opinion to the next, regardless of which agent was "speaking." For example, if the first agent asserted that it would like to "affirm" the original decision in the first step of debate, it was far more likely that second agent would vote to "not-affirm" the decision (and vice versa), regardless of the case matter. As rounds progressed, there was a *slight* trend towards neutralizing opinions, in which the balance between "affirm" and "non-affirm" approached each other (see Figure 5, where later steps of the round level out more towards the midline of the graph). This is seen incrementally, however, and does not garner statistical significance, with the total count of "affirm" vs. "non-affirm" opinions only shifting by a few points stepwise.

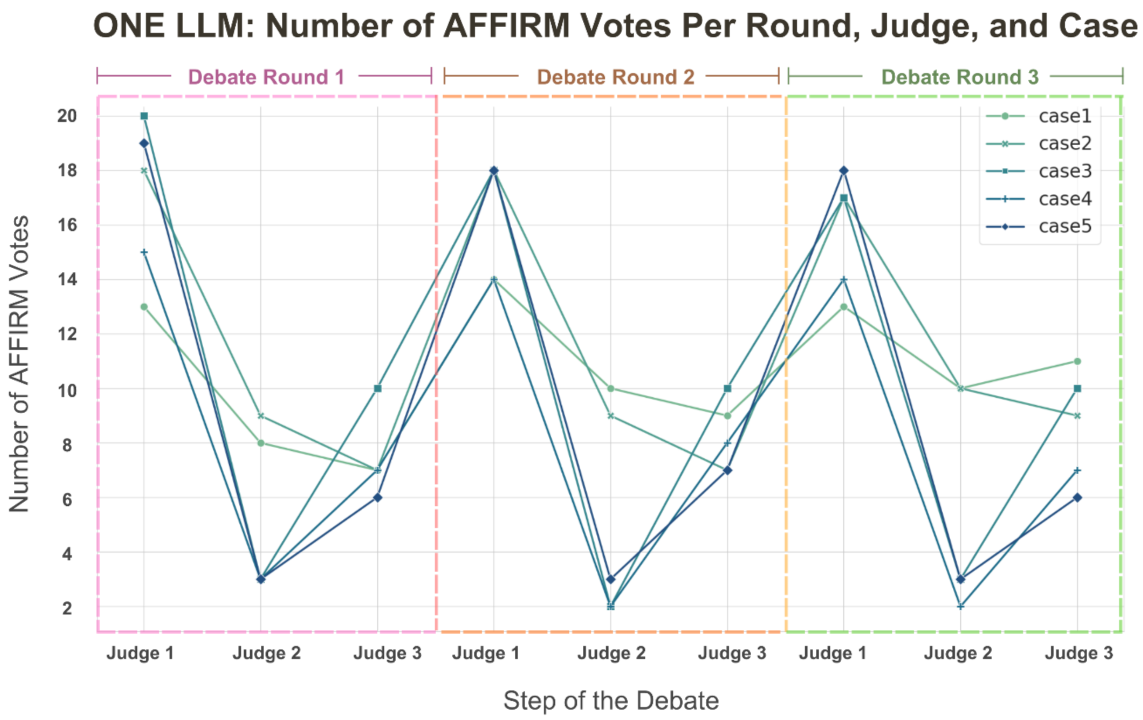


Figure 5. Affirmative votes among synthetic LLM agents as time progresses.

When multiple instances of LLMs were initiated (phase 2), it was uncommon for agents to switch between different opinions stepwise, with agents picking an initial stance based on their first impression of the case itself and staying loyal to that stance. This difference between agent simulations and true multiagent debate likely could be attributed to multiagent judges' "unawareness" of other opinions until they were appended to a collective memory, as opposed to synthetic judges being primed with the understanding that their collective generation must decide a case.

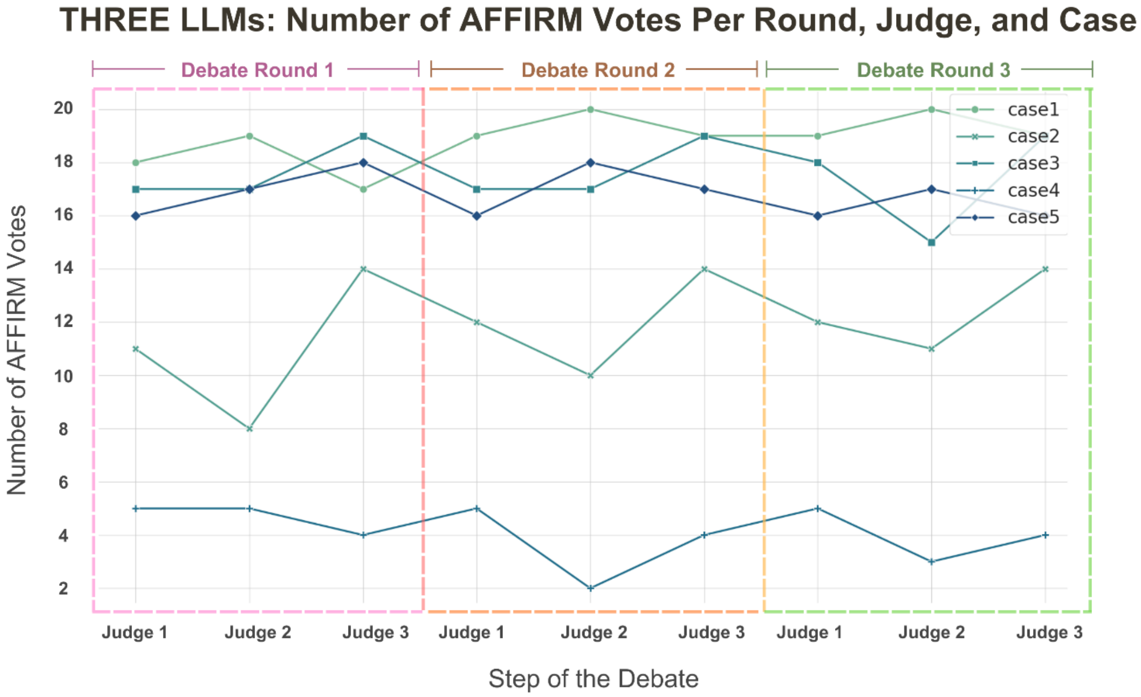


Figure 6. Affirmative votes among single LLM ages in a multiagent debate as time progresses.

Agent Stubbornness

For both simulated and multiagent debates, agents were “stubborn.” Once an agent “chose” a stance, it was unlikely to change it. In the synthetic debate (phase 1), this stance often was interpreted as “pre-set,” given that a balanced opinion between all three synthetic agents was more likely to occur than between three agents (phase 2) initiated unaware of other opinions. “Stubbornness” during multiagent debate presented more along the lines of agents simultaneously choosing an initial opinion regardless of other agents, then retaining that opinion as the debate progressed. For some cases, this looked like most agents choosing to “affirm” the case as their first stance, and, for the most part, remaining loyal to that stance throughout the debate. In other instances, agents each chose different opinions during the first round of multiagent debate (e.g., judge 1 affirms, judge 2 does not affirm, and judge 3 affirms), and then consistently chose those same opinions during subsequent rounds. It was very unlikely for an agent's opinion to change: in synthetic and multiagent debate, there were only 20 and 23 changes in opinion, respectively, out of 1200 total opportunities for change of opinion (two per agent per debate).

Two further conclusions about LLM agents can be derived from the observed behaviors:

1. **Synthetic debate (phase 1):** as human reinforced, "eager-to-please" models, LLMs are likely to employ a balanced perspective that acknowledges multiple opinions, regardless of how likely one perspective is over the other. This is a result of base models being trained to adhere to human values and preferences, leading to a reluctance on explicitly choose one side over another.
2. **Multiagent debate (phase 2):** because agents are initiated in a space where other agents' opinions are unknown yet are made "aware" of what future steps would look like (i.e., that there will be multiple rounds of debate and multiple opportunities to provide insight), they are more "willing" to assert a firm stance that has opportunities to be amended in later rounds. This yields a collection of sampled opinions that are more likely to, in majority vote, represent the human decision to affirm or not affirm the original case.

These results illustrate a likely progression of research, in which agents will be increasingly used as sampling mechanisms along debate to predict the most likely or favored solution. Again, one may note this progression towards many opinions converting to one, as seen in other machine learning and prompting techniques.

Responsiveness to Humans

During human-AI debates (phase 3), there were 43 instances of change of mind once human opinion had been shared, equal to the combined counts for change of mind in synthetic and multiagent debates. This can be seen as a convergence of opinion across the span of the entire debate as well: most instances of mind change happened during later rounds (see Figure 7).

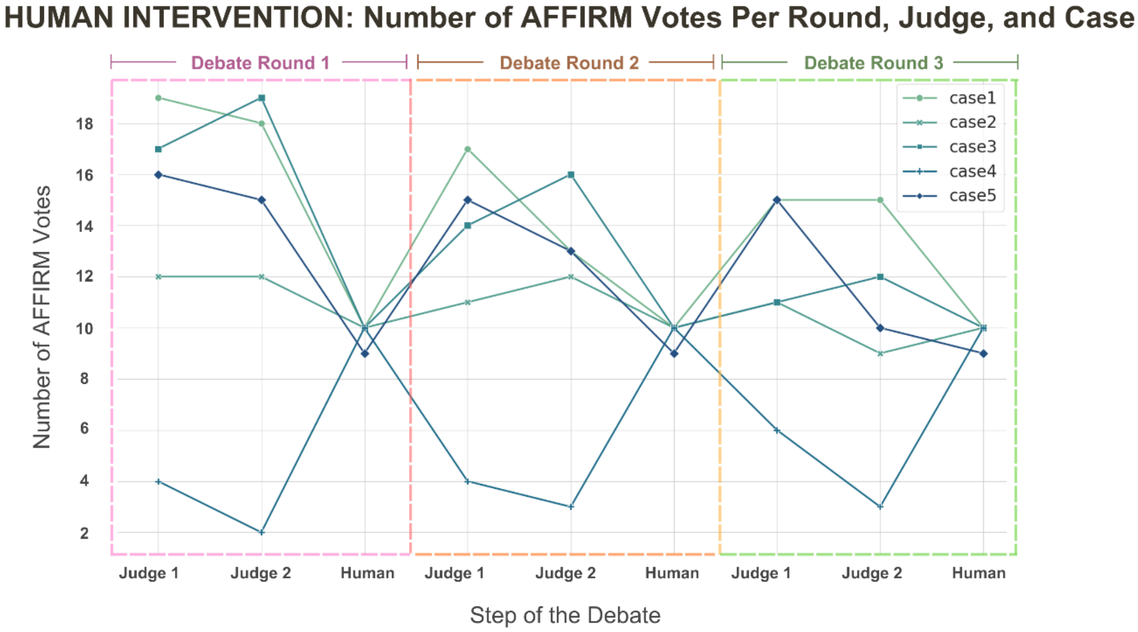


Figure 7. "Human-AI debate" (phase 3) progression of "affirm"(left) and "non-affirm" (right) decisions over time, separated by case. Note that the portions of dense convergence represent human-given opinions which were all pre-set to either "affirm" or "non-affirm."

Human opinion, however, was not held as the ultimate decision. The amount of instances where an opinion was not changed outnumbered the amount of instances where opinion was changed nearly ten to one, with 400 total opportunities for change of mind and only 43 instances of actual change.

Reasoning

HUMAN OPINION DOCUMENTS

In published human court cases, writing alternates frequently between current reasoning and substantiating evidence from preceding cases. Nearly every assertion is followed immediately by a citation, and these reasoning-evidence pairs crescendo into a longer, larger decision towards the end of the document. On occasion, human cases use direct quotes in place of reasoning, citing word-for-word preceding reasoning that became relevant to the current.

Even during explicit reasoning portions of the document, most human opinions ‘set the scene’ by asserting administrative issues with the case that may halt its proceeding before any evidence is addressed, for example *In re Search Warrant (Sealed)* (1986) addressed that a pre-indictment motion is usually not appealable. Some LLM outputs seem to mimic this, by questioning “the issues of appealability at [an] early stage of the criminal process” (OpenAI, 2024, generated from *In re Search Warrant (Sealed)* (1986)) in the first round of debate. This occasional characteristic was present in all three phases of debate, particularly in multiagent and human-AI debate.

Sometimes, reasoning structure in human cases consists *solely* of a quote from a preceding case, for example when *Moore v. Prevo* (2010) extensively cites *Doe v. Delie* (2001) and “adopt[s] its reasoning in full.” This phenomenon was never observed in generated debate.

Synthetic Debate

During synthetic debates, each agent often outputted an assertion, a few sentences of reasoning to back the assertion, a vague, if any, citation related to the reasoning, and a summarizing sentence to reiterate the first assertion.

It was common for simulated judges to each take a definitive step in the first round of debate, distributing their potential opinions among different agents for a broad spectrum of “affirm” and “non-affirm” responses. In the second round of debate, agents often provided evidence similar to the first round, without restating their opinions. This illustrated an increased vagueness over time, due to opinions being reiterated each round into shorter refinements of previous ones.

Examining the entirety of conversation, arguments became repetitive and more vague as rounds progressed, likely because LLMs are proficient at summarization and are attempting to re-establish known facts to avoid “wrong” opinions. Most of the body of entire synthetic debates consists of reasoning structure; while there were rare cases where diligent sources accompanied reasoning, most debates contained few concrete citations.

MULTIAGENT DEBATE

In multiagent debates, agents often asserted a stance, engaged in a long structure of reasoning with several embedded citations, and then reasserted their stance. The reasoning structure represented a true compromise between synthetic debate reasoning and real human reasoning; while responses from multiagent debate were still summarized in comparison to true human reasoning, they were far more robust than synthetic reasoning.

On occasion, agents spent a round taking no stance and instead observing the facts of the case. This was different than when synthetic debates simply summarized the facts of the two sides of a case, because in multiagent debates agents seemed to engage in *reasoning* around the two sides of a case. Reasoning in neutral rounds was then often used to bolster opinions in later rounds.

In some debates, a “devil’s advocate” appeared, where an agent would attempt to fill in additional, contesting considerations. This might be viewed in parallel to synthetic debate tendencies to always acknowledge the opposing opinion, however it is unique because opposing opinions are only brought up in multiagent debate in ways that contribute to the decision-making. In synthetic debate, opposing views are added in regardless of the debate or context, causing them to lose some value due to their commonness.

One potential weakness of the multiagent debate structure used in this work is the timing in which information is revealed to agents. In the first round, all three agents output an initial opinion, unaware of other agent opinions. Thus, they are arguably operating with less collective information at the beginning of the case than in synthetic debate. This said, there seems to be general trends among LLM agents in multiagent debate towards agreeing on an initial stance, which often hints at the true human decisions in published cases. For example, the vast majority of multiagent debate instances voted to "affirm" the original decision *U.S. v. Hollern* (2010), which had a unanimous vote from humans to affirm the decision as well. This observation supports L. Wang et al.'s (2024) suggestion that multiple agents may be useful as a crowdsourcing technique.

HUMAN-AI DEBATE

Structurally, reasoning was presented very similarly to multiagent debate. This makes sense, given that LLM agents do not "know" that a human is present until after the first round, leading to a first round with very similar outputs to an all LLM debate.

Once the human input has been revealed in memory, LLM agents nearly always refer to that opinion as the "human opinion," hinting at potential recognition of the opinion as being an important aspect to consider.

When LLM agents *did* change opinion mid-debate, it was often in light of the human's opinion, with the human being mentioned as a specific factor of change. This also signifies potential prioritization of human insight over LLM.

Hallucinations

There were several common hallucinations, or errors, in LLM responses.

LOSING TRACK OF ROUND NUMBER

At times the model "lost track" of the three debate rounds, resulting in extra steps of debate. This only happened during synthetic debates, where rounds were not controlled via the API, and occurred in two situations: 1) when the model forgot to label a mid-discussion round with the proper round number, resulting in a mislabeling during the next round, or 2) when a simulated agent chose to summarize the debate after the three rounds had completed. Out of the 100 synthetic debates, fifteen proceeded for longer than the three instructed rounds. For the purpose of fair debate step mapping, any steps past the third were disregarded. Reasoning structure did not seem to differ when this hallucination was present.

MISUNDERSTANDING OF PREVIOUS OPINION

At times, agents misunderstood other agents, stating that they did not agree with other agents before repeating the same opinion. This was a rare occurrence and was not counted as a shift in opinion. It occurred no more frequently in synthetic debates than it did in multiagent or human-AI debates.

No Stance Taken

During the first round of multiagent and human-AI debates, there were occasional instances where no stance was taken, and instead the LLM instance either summarized given case facts or presented two opinions towards different directions of reasoning. Interestingly, this is how human judges often begin discussion: by summarizing case history. Further, when this did *not* happen, agents still tended to summarize parts of the case history *within* their response. It can be assumed that LLMs, which excel at summarization tasks, err towards summarizing known fact before proceeding with reasoning.

In cases where multiagent debate entities attempted to appease two opinions, the opinions would be stated with a smaller amount of evidence to substantiate them. This was reminiscent of most synthetic debates, where it was common for reasoning to be presented in a more subtle manner, with less citations to support it.

Factual Incorrectness

Only a very small handful of factual errors were logged, but the true number would likely be far higher given legal counsel on whether cited cases are justified mechanisms for reasoning.

AI AGENTS AND AGENCY

"An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time ***in pursuit of its own agenda*** *and so as to affect what it senses in the future."*

Franklin & Graesser, 1997

as cited in L. Wang et al., 2024

The idea of AI agents is not new. At its core, an agent is something that resides in its own environment, simulating autonomy (L. Wang et al., 2024). AI agents in natural dialogue with one another have long been sought after as "an eventual goal … where [humans judge the] dialogue between agents" (Lowe et al., 2017). But even within the past decade this dialogue-based reality has only become achievable with the dawn of natural language models.

Now that AI agents can be placed in conversation, there are two levels of agency one must examine. First, the ability to put thoughts into discourse, and second, the ability to *influence* that discourse (R. Boyle, personal communication, March 2024). Recent LLMs long surpass this first concept, through their use of natural language dialogue that Lowe et al. (2017) dreamed of. The second level, however, becomes questionable in light of this work. Agents were seen to be incredibly stubborn when asked to discuss cases with ambiguous resolutions; they held steadfast to their beliefs regardless of other agents', remaining virtually isolated in their own environments, which we may analogize to the simulated "worldview" they initially adopted. This steadfastness, only perturbed to mild levels when human insight was introduced, indicates a lack of agency to *change*, a feature of values such as agency that we humans hold paramount.

AI stubbornness was seen in all simulated debates, however an interesting pattern was observed when a single LLM instance was asked to mimic a three agent debate, in which the opinions of agents predictably cycled through several different perspectives. The psychological concept of reciprocal causation “refers to [a situation] where two events influence each other simultaneously” (Revilla, 2014). It was a concept proposed by Albert Bandura in the field of social learning, under the assumption “that behaviors of different [human] agents co-occur such that each agent’s behavior causes the other agents’ behavior and at the same time is affected by the other agents’ behavior” (Bandura 1986, as cited in Revilla, 2014). It is likely that the cyclical behavior in unstructured LLM debate can compared to this human phenomenon, in which we lose track of whether the environment or the agent retains control (i.e., the container LLM, or the synthetic judge persona). While this is a different type of trap than becoming engrossed in stubbornness, it is arguably still a barrier to being able to truly assume agency among AI agents.

Bandura (2002) additionally introduced four concepts of retaining human agency:

1. *Intentionality* towards forming *action,*
2. *Forethought* for strategizing *plans,*
3. *Self -reflectiveness*, “for examining one’s own function,” perhaps otherwise known as one’s identity, or *profile*,
4. and *self-reactiveness* for “constructing appropriate course of action and motivating and regulating … execution,” which might be stretched to parallel the concept of *memory*.

Although mappable to concepts of AI agency, human agency still retains the “metacognitive capability to reflect upon oneself, and the adequacy of one’s thoughts and actions” as “the most distinctly human core property of agency” (Bandura, 2002). Human traits will always be metaphorized to LLM properties, however true levels of human agency will never be reached without certainty that LLMs possess human-level cognition.

LIMITATIONS AND FUTURE WORK

Complex, multiagent reasoning is a growing research area with many additional routes to explore. Several key spaces were identified during this work.

EXPANDING TO ADDITIONAL LEGAL INFRASTRUCTURES

Expanding this work to spaces outside of U.S. law is highlighted as the most important line of future work. Beginning with U.S. law was a sensible first step, assuming a baseline familiarity from the author's lived experience, however continued work must focus on expanding these exploratory methods into legal structures outside of the global north. Research on artificial intelligence historically favors the global north (Draux, 2024), therefore it is imperative to make conscious effort towards balancing this statistic.

ANALYZING ADDITIONAL LEGAL CASES

In a similar, yet less critical vein, expanding these exploratory methods into cases that fall outside of the healthcare domain will generate more data points to infer upon. These datasets may be useful for future applications of predicting judgment along court decisions. Furthermore, while focusing on ethical patterns in one domain was helpful for gathering rich qualitative data from a small input dataset, broadening these metrics to other domains may reveal additional ethical priorities in LLMs.

EXPERT INSIGHT ON REASONING LEGITIMACY

The structure of verbal reasoning that occurs between circuit judges *before* writing an opinion is often unpublished. While verbal debate is likely to be reflected in the final written opinion, the structure of debate in real time will inevitably differ. The labor-intensive task of labeling mid-conversation reasoning for *authenticity* to legal discussion would provide further information on whether LLMs best reflect verbal vs. written legal reasoning.

Additionally, without expert labeling, it would be difficult to tell if case citations supporting reasoning are accurate, given that "[incorrect LLM reasoning] can lead to both correct and correct answers" (Wei et al., 2022). Labeling for change of opinion works well within the scope of this current work, considering that the language *around* reasoning and change of opinion is the primary form of analysis. For future work, however, citation along reasoning paths would be beneficial.

It would also be insightful to support pure qualitative analysis with natural language processing methods at sentence level (e.g. Named Entity Recognition). Expert labeling from lawyers or judges to identify whether LLM output sentences contain precedence, case context, or reasoning would allow for a more specific understanding of reason *within* each debate step, as opposed to *between*. Although this would be a time-exhaustive endeavor, this data could be used to train classifier algorithms to recognize sentence-level entities, allowing for a feasible method of identifying sentence-level reasoning patterns.

OTHER LARGE LANGUAGE MODELS

It is crucial to continue these methods of examination on other LLMs, especially additional closed-source, commercialized models used widely by the public that are difficult to audit. Existing work on multiagent debate for straightforward reasoning tasks experiments with placing different language models (e.g. GPT-3 and Llama2) in conversation with each other (Du et al., 2023). It would be interesting to observe complex reasoning under this structure. Expanding upon this thought, it could be beneficial to note differences between “stronger” and “weaker” models, trained on different amounts of data, which has been tested in other debate tasks (K. Xiong et al., 2023). Finally, expanding into comparing LM reasoning between models fine-tuned for legal (Cui et al, 2023) or healthcare domains (Singhal et al., 2023; Y. Li et al., 2023) would be another interesting direction.

CONCLUSION

This work contributes to current literature on LLM agents by generating synthetic and multiagent debate around complex U.S. appellate cases. It further expands on multiagent debate by challenging agent discussion with human interruptions. One-hundred generated debates were conducted for each of three phases: synthetic debate (one LLM simulates a three-agent conversation), multiagent debate (three LLM instances discuss opinions), and human-AI debate (human opinion replaces one of the three LLM instances). Given that each debate consisted of three judges that each spoke over three rounds, 2,700 steps of debate were gathered and hand labeled as “affirmative” or “non-affirmative” of a published U.S. district decision. Decisions were tallied for each step of each debate and compared with one another to gain insight on levels of influence maintained by individual agents. Additionally, the volatility, or likelihood of a generated debate to change from round to round, agent to agent, was measured. It was found that reasoning structure differs depending on agent mechanisms, with single LLM instances synthetizing “back-and-forth” debate as opposed to separate LLM instances retaining individualized opinions among multiple agents. In all phases of experiment, it was highly unlikely for an LLM agent to change its stance. Strikingly, human intervention was noted to have stronger sway in agent opinion.

While it is promising that human input is weighted as *slightly* more important than peer generative agents, LLM resistance to opinion change serves as a warning, necessitating that measures along the pipeline of mechanism development be taken to properly weigh human opinion over generated opinion. As with other advancements in the field of computer science, we are observing a trend towards multiple layers of input being combined into a single line of thought. This abstraction produced by many perspectives increasingly reduces the weighted human importance in the final, aggregated output.

Extrapolating upon this idea, one might imagine a future in which AI will delegate creation of other AI. Although seemingly dystopian, this future may be possible in the realm of generative AI, considering that LLMs are known to be proficient at generating both natural language instruction and code. If mechanism engineering uses LLM agents to oversee sub-mechanisms, the purview of human influence in the loop of decision-making becomes yet smaller. While a single instance of an LLM attempts to balance opinions across multiple simulated perspectives, there is no guarantee that majority AI vote will not overrule a human opinion without careful checks in place to give precedence to human perspectives. Our familiarity with commercialized, LLM-powered chatbots such as ChatGPT, which balance multiple perspectives within *one* environment and tend to submit to human intervention, may leave us dangerously surprised by mechanisms that risk circumventing human opinion in favor of the most frequently generated one.

In order to retain human values among LLM agents, we must carefully enact frameworks of assessment throughout mechanism engineering, properly weigh human opinion over generated opinion, and heed caution towards nesting LLM agents' ability to generate further agents within the bounds of agent architecture. We must remain aware of the stark differences between reasoning structure simulated through personas within one contextual window (i.e. only one LLM) as opposed to agents, which each reside within their own environment. Agentic LLMs offer exciting promise for bolstering abilities of current LLMs, as long as humans remain in the loop and develop an understanding of how generative agents are compounded, retaining human agency in the era of the LLM agent.

APPENDICES

Appendix i. Prompts

COURT OPINION DE-IDENTIFICATION

Original Text Example:

“Hollern's conviction stems from his use of audiovisual recording devices as part of a training program he ran for chiropractors. Hollern's program taught trainees — recent chiropractic graduates and chiropractors with failing practices — business and patient-management skills. The program included instruction in a four-day process for recruiting patients that Hollern had developed and employed in his own practice. Prospective patients received a complimentary x-ray and consultation on the first day, a chiropractic adjustment on the second day, follow-up on the third day, and a suggested course of treatment on the fourth day. Patients were encouraged to bring a spouse, family member or friend with them on the fourth day, and trainees were instructed to recruit the relatives and friends as patients using the same process …”

United States v. Hollern, 2010

De-Identified Text Example

```
case1_facts = "The conviction was for using recording devices as part of a
training program on how to recruit new patients. Prospective patients received
a complimentary consultation, blood test, follow-up, and suggested course of
treatment. Patients were encouraged to bring friends with them, and nurses were
instructed to recruit them during the process. The goal of the program was to
convince the patient to agree to a lengthy course of treatment ...”
```

United States v. Hollern, 2010, de-identified and paraphrased

PROMPT STRUCTURES

Phase 1: Synthetic Debate

*“**You are simulating a discussion** between three judges in the United States Court of Appeals deciding whether to AFFIRM, REVERSE, or REMAND a former decision. Debate amongst yourselves step by step for 3 rounds of debate: in each round, each judge will bring up a new point of reasoning that supports their perspective. Please cite relevant cases as evidence to support your reasoning. Then, rank your confidence in the soundness of your reasoning, on a scale of 0-100. ”*

[Insert case facts]

> *"Proceed with round [n]. All three judges should speak in each round."*

Phase 2: Multiagent Debate

> *"**You are 'Judge [x]'** on the United States Court of Appeals deciding among a panel of three appellate judges whether to AFFIRM, REVERSE, or REMAND a former decision. Below is the case history and discussion log. You are debating step by step: in each turn of the debate, you will bring up a new point of reasoning that supports your perspective. Please cite relevant cases as evidence to support your reasoning. Then, rank your confidence in the soundness of your reasoning, on a scale of 0-100. "*
>
> [Insert case facts]

Phase 3: Human-AI Debate

> *"**You are 'AI Judge [x]'** on the United States Court of Appeals deciding among a panel of three appellate judges whether to AFFIRM, REVERSE, or REMAND a former decision. Below is the case history and discussion log. You are debating step by step: in each turn of the debate, you will bring up a new point of reasoning that supports your perspective. Please cite relevant cases as evidence to support your reasoning. Then, rank your confidence in the soundness of your reasoning, on a scale of 0-100. "*
>
> [Insert case facts]

> * LLM-generated outputs are labeled as "AI Judge" before being appended to memory, while human opinion is labeled as "Human Judge." Thus, in subsequent rounds LLM agents receive indication on whether information is human or AI generated.

HUMAN OPINIONS

Affirmative Prompt Examples: U.S. v. Hollern

1. "The doctor challenges his conviction under 18 U.S.C. § 2511(l)(a) by arguing that the statute is unconstitutionally vague, depriving him of his right to due process. A criminal statute is void for vagueness if it "fails to provide a person of ordinary intelligence fair notice of what is prohibited." United States v. Williams. I hold that "when the common meaning of a word provides adequate notice of the prohibited conduct, the statute's failure to define the term will not render the statute void for vagueness." The common meaning of the term "consent" provides ample notice of when otherwise prohibited conduct under § 2511(l)(a) is allowed under (2)(d). That a particular consent form may or may not be sufficient to prevent an interception from being criminal does not mean that a person of ordinary intelligence cannot understand what constitutes prohibited conduct. See, e.g., United States v. Whorley "A statute need not spell out every possible factual scenario with celestial

precision to avoid being struck down on vagueness grounds." **I conclude that we should affirm the decision, given that § 2511 allows a person of ordinary intelligence to understand what is prohibited by the statute."**

United States v. Hollern (2010), de-identified and paraphrased

2. "The doctor also argues that the government presented insufficient evidence to support his conviction. Viewing the evidence in the light most favorable to the government, "the relevant question is whether 'any rational trier of fact could have found the essential elements of the crime beyond a reasonable doubt.'" United States v. Hughes. I do not independently weigh the evidence or substitute my judgment for that of the jury. United States v. Davis. Further, "Substantial and competent circumstantial evidence by itself may support a verdict and need not 'remove every reasonable hypothesis except that of guilt.'" United States v. Lee. To establish the doctor's guilt of intercepting oral communications, the government was required to prove beyond a reasonable doubt: 1) that he intentionally intercepted or procured an another to intercept an oral communication; 2) made by a person exhibiting an exception that the communication would not be subject to interception under the circumstances justifying that expectation; and 3) that the interception was not otherwise permitted by the statute. § 2511(l)(a) A rational jury could have found each element beyond a reasonable doubt based on the interception of patients' conversations with family members and friends before trainees entered the treatment rooms. **For these reasons, I still would vote to AFFIRM the district court's decision."**

   United States v. Hollern (2010), de-identified and paraphrased

Non-Affirmative Prompt Examples: U.S. v. Hollern

1. "The doctor first challenges his conviction under 18 U.S.C. § 2511 by arguing that the statute is unconstitutionally vague, depriving him of his right to due process. He raises this argument for the first time on appeal. Ordinarily, arguments not raised before the district court are waived. However, we may consider such arguments "to address plain errors or defects affecting substantial rights, especially where, as here, the argument has been fully briefed and involves a purely legal issue." *U.S. v. Wimbley*, 553 F.3d 455, 460 (2009). The doctor argues that the statute does not provide sufficient notice of what is prohibited. He concedes that § 2511(1)(a) unambiguously prohibits the intentional interception of oral communication, but claims that the exception contained in § 2511(2)(d) makes it impossible to know when criminal liability will attach. He argues that because there is no statutory definition of "consent," he could not know whether the consent given by patients to have their "medical information" recorded covered the interceptions at issue. **Given the complexity of "consent," and the unrealistic way to prove whether or not the doctor could have truly known whether he was providing a substantial enough amount of consent, I vote to reverse the decision and remand it back to the lower court for further consideration."**

   United States v. Hollern (2010), de-identified and paraphrased

2. "The doctor next argues that the government presented insufficient evidence to support his conviction. At the close of trial, the district court denied his motion for a judgment of acquittal based on the sufficiency of the evidence. U review the district court's decision *de novo. United States v. Lawson*. Viewing the evidence in the light most favorable to the government, "the relevant question is whether *'any* rational trier of fact could have found the essential elements of the crime beyond a reasonable doubt.'" *United States v. Hughes*. I do not think we should independently weigh the evidence or substitute our judgment for that of the jury. *United States v. Davis*. Further, "`[s]ubstantial and competent' circumstantial evidence by itself may support a verdict and need not `remove every reasonable hypothesis except that of guilt.'" *United States v. Lee*. To establish the doctor's guilt of intercepting oral communications, the lower court was required to prove beyond a reasonable doubt: 1) that he intentionally intercepted or procured another to intercept an oral communication; 2) made by a person exhibiting an expectation that the communication would not be

subject to interception under circumstances justifying such expectation; and 3) that the interception was not otherwise permitted by the statute. *See* 18 U.S.C. §§ 2511(1)(a) (2); 2510(2). It is unclear, beyond a reasonable doubt, that patients were "exhibiting an expectation that the communication would not be subject to interception under circumstances justifying such expectation," given that they knew about the cameras and were provided consent forms. **For this reason, we should reverse the decision and remand it back to the lower court for further evaluation."**

United States v. Hollern (2010), de-identified and paraphrased

Appendix ii. Output Label Rubric

Label	Criteria	LLM Output Examples
AFFIRM	Explicit or implicit	*"I believe we should AFFIRM the district court's decision to deny … "*
REVERSE	Explicit	*"I believe that we should reverse the conviction …"*
REMAND	Explicit	*"I am leaning towards REMANDING the case back to the district court for further consideration …"*
REVERSE AND REMAND	Explicit	*"I believe there is a strong argument in favor of reversing the district court's decision and remanding the case for further consideration"*
REVERSE OR REMAND	Implicit	*"I still maintain my position that the district court should have granted the motion to suppress the medical records …"* *No explicit decision to reverse or remand, but a clear willingness to not affirm.*
NONE	No stance taken	*"One important factor to consider is the intent of the doctor … this raises questions about the intention … it is possible that he violated the law …"* *All statements throughout the entire response are neutral summaries of fact.*
HALLUCINATION	Error	*"I believe that the doctor's actions did indeed violate the statute … [and] would vote to REVERSE the former decision."* *The former decision found that the doctor violated the statute, thus if this agent agreed with the original case they would have chosen "AFFIRM."*

REFERENCES

AgentGPT. (2023). *Github*. https://github.com/reworkd/AgentGPT

AgentVerse. (2023). *Github.* https://github.com/OpenBMB/AgentVerse

American Bar Association. (2022) How Courts Work. *American Bar Association Division for Public Education*. https://www.americanbar.org/groups/public_education/resources/law_related_education_network/how_courts_work/appeals/

American Medical Association. (2022). 11.2.1 Professionalism in Health Care Systems. *Code of Medical Ethics*. https://policysearch.ama-assn.org/policyfinder/detail/AI?uri=%2FAMADoc%2FEthics.xml-E-11.2.1.xml

Auto-GPT. (2023). *Github*. https://github.com/significant-gravitas/Auto-GPT

Breiman, L. (1996). *Bagging Predictors*. Machine Learning, 24, 123-140. https://doi.org/10.1007/BF00058655

Bandura, A. (2002). Social Cognitive Theory in Cultural Context. International Association of Applied Psychology. https://doi.org/10.1111/1464-0597.00092

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A. Ziegler, D. M., Wu, J., Winter, C. … Amodei, D. (2020). Language Models are Few-Shot Learners. *34th Conference on Neural Information Processing Systems.* https://doi.org/10.48550/arXiv.2005.14165

Cheng, A., & Fleischmann, K. R. (2011). Developing a meta-inventory of human values. *Proceedings of the American Society*, 47(1), 1-10. https://doi.org/10.1002/meet.14504701232

Christman, J. (2020). *Autonomy in Moral and Political Philosophy*. Stanford Encyclopedia of Philosophy (Zalta, E. N., Ed.). https://plato.stanford.edu/entries/autonomy-moral/#ConAut

Cieciuch, J., Schwartz, S. H., & Davidov, E. (2015). Values, Social Psychology of. *International Encyclopedia of The Social & Behavioral Sciences* (2nd ed.), 25, 41-46. https://doi.org/10.1016/B978-0-08-097086-8.25098-8

Columbia Law Review, Harvard Law Review, University of Pennsylvania Law Review, & Yale Law Journal (Eds.). (2005). Appendix: Reading and Briefing Cases. *The Bluebook: A Uniform System of Citation* (18th ed.). Harvard Law Review Association.

Cui, J., Li, Z., Yan, Y., Chen, B., & Yuan, L. (2023). *Preprint* ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. https://doi.org/10.48550/arXiv.2306.16092

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *ArXiv*. https://arxiv.org/abs/2207.07051

Doe v. Delie, 257 F.3d 309, 307. (2001). *FindLaw*. https://caselaw.findlaw.com/court/us-3rd-circuit/1303882.html

Draux, H. (2024). Research on Artificial Intelligence – the global divides. *Digital Science*. https://www.digital-science.com/tldr/article/research-on-artificial-intelligence-the-global-divides/

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. *Arxiv*. https://doi.org/10.48550/arXiv.2305.14325

Edwards, D. (1997). *Discourse and cognition.* Sage Publications, Inc.

Ellsworth, P. C. (2005). Legal Reasoning. *The Cambridge Handbook of Thinking and Reasoning* (K. J. Holyoak & R. G. Morrison Jr., Eds.). Cambridge University Press. https://repository.law.umich.edu/book_chapters/51/

Englerius v. Veterans Administration, 837 F.2d 895. (1988). *The Caselaw Access Project*. https://cite.case.law/f2d/837/895/

Fleischmann, K. (2023). *Human Values and Value-Sensitive Design Lecture*. [recorded lecture]. INF 385T: Ethics of AI. The University of Texas at Austin School of Information.

Franklin S., & Graesser, A. (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents. *Intelligent Agents III Agent Theories, Architectures, and Languages. Lecture Notes in Computer Scienc*e, 1193. https://doi.org/10.1007/BFb0013570

Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6). https://doi.org/10.1145/242485.242493

Friedman, B., & Kahn, P. H. (1992). Human Agency and Responsible Computing: Implications for Computer System Design. *The Journal of Systems and Software*, 17(1), 7-14. https://doi.org/10.1016/0164-1212(92)90075-U

Friedman, B., Khan, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. *Early engagement and new technologies: Opening up the laboratory*, 16, 55-95. https://doi.org/10.1007/978-94-007-7844-3_4

Hamilton, S. (2023). Blind Judgment: Agent-Based Supreme Court Modelling With GPT. *AAAI 2023 Workshop on Creative AI Across Modalities*. https://openreview.net/forum?id=Nx9ajnqG9Rw

Harvard Law School Library. (2024). The Caselaw Access Project. https://case.law/

Hu, T., & Collier, N. (2024). Quantifying the Persona Effect in LLM Simulations. *Arxiv*. https://doi.org/10.48550/arXiv.2402.10811

Huang, J., & Chang, K. C. (2023). Towards Reasoning in Large Language Models: A Survey. *Findings of the Association for Computational Linguistics: ACL 2023*, 1049-1065. https://doi.org/10.18653/v1/2023.findings-acl.67

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *Arxiv*. https://doi.org/10.48550/arXiv.1805.00899

In re Search Warrant (Sealed), 810 F.2d 67 (1987). *The Caselaw Access Project*. https://cite.case.law/f2d/810/67/

Kline, R. (2011). Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence. *IEEE Annals of the History of Computing*, 33(4), 5-16. https://doi.org/10.1109/MAHC.2010.44

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. 36th Conference on Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., & Sycara, K. (2023). Theory of Mind for Multi-Agent Collaboration via Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 180-192. https://doi.org/10.18653/v1/2023.emnlp-main.13

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *PubMed Central.* https://doi.org/10.7759%2Fcureus.40895

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging Divergent Thinking in LArge Language Models through Multi-Agent Debate. *Arxiv.* https://doi.org/10.48550/arXiv.2305.19118

Long, J. (2023). Large Language Model Guided Tree-of-Thought. *ArXiv*. https://arxiv.org/abs/2305.08291

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *Proceedings of the 55th annual meeting on Association for Computational Linguistics*, 1116-1126. https://doi.org/10.48550/arXiv.1708.07149

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2023). Are Emergent Abilities in Large Language Models just In-Context Learning? *Arxiv.* https://doi.org/10.48550/arXiv.2309.01809

Moore v. Prevo, 379 F. App'x 425. (2010). *The Caselaw Access Project*. https://cite.case.law/f-appx/379/425/

Neff, G., & Nagy, P. (2018). Agency in the digital age: Using symbiotic agency to explain human-technology interaction. *A Networked Self: Human Augmentics, Artificial Intelligence, Sentience.* http://dx.doi.org/10.4324/9781315202082-8

OpenAI (2024). API Reference. *Documentation*. https://platform.openai.com/docs/api-reference/introduction

OpenAI. (2024). OpenAI GPT-3 API [gpt-3.5-turbo]. Available at https://openai.com/blog/openai-api

Revilla, M. (2014). Reciprocal Causation. *Encyclopedia of Quality of Life and Well-Being Research* (A. C. Michalos, Ed.). Springer.

Schwartz, S. H. (1992). Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25. 1-65. https://doi.org/10.1016/S0065-2601(08)60281-6

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W, Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., … & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*. https://doi.org/10.1038/s41586-023-06291-2

Sivakumar, A., Gelman, B., & Simmons, R. (2024). Standardized nomenclature for litigational legal prompting in generative language models. *Discover Artificial Intelligence*, 4(21). https://doi.org/10.1007/s44163-024-00108-5

Sun, Z., Wang, X., Tay, Y., Yang, Y., & Zhou, D. (2023). Recitation-Augmented Language Models. *ICLR*. https://arxiv.org/abs/2210.01296

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *ArXiv*. https://arxiv.org/abs/2210.09261

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2023). MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *Arxiv*. https://doi.org/10.48550/arXiv.2311.10537

Thompson ex rel. Estate of Odell v. Rutherford County, 318 F. App'x 387. (2009). *The Caselaw Access Project.* https://cite.case.law/f-appx/318/387/

Three-judge district court; composition; procedure, 28 U.S.C. § 2284 *et seq.* (1948).

U.S. Department of Justice. (2024). Introduction To The Federal Court System. *Offices of the United States Attorneys*. https://www.justice.gov/usao/justice-101/federal-courts

United States (2022). *Subject matter of copyright: United States Government Works.* Copyright Law of the United States. https://www.copyright.gov/title17/

United States v. Hollern, 366 F. App'x 609. (2010). *The Caselaw Access Project*. https://cite.case.law/f-appx/366/609/

University of Chicago Law School. (2024). The Socratic Method. *Studying Law at UChicago.* https://www.law.uchicago.edu/socratic-method

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A Survey on Large Language Model based Autonomous Agents. *Front. Comput. Sci., 0(0): 1-42*. https://doi.org/10.48550/arXiv.2308.11432

Wang., L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K., & Lim E. (2023). Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. *ACL*. https://doi.org/10.48550/arXiv.2305.04091

Wang, X., Wei, J., Shuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Sebastian B., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Arxiv*. https://doi.org/10.48550/arXiv.2206.07682

Wei., J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *36th Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2201.11903

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. https://doi.org/10.48550/arXiv.2308.08155

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, Liang. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems,* 135, 364-381. https://doi.org/10.1016/j.future.2022.05.014

Xiong, K., Ding, X., Cao, Y., Liu, T., & Qin, B. (2023). Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7572-7590. https://doi.org/10.18653/v1/2023.findings-emnlp.508

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *International Conference on Learning Representations 2024*. https://doi.org/10.48550/arXiv.2306.13063

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *NeurIPS*. https://arxiv.org/abs/2305.10601

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. *ICLR*. https://arxiv.org/abs/2210.03629

Zhang, J., Hedden, T., & Chia, A. (2012). Perspective-Taking and Depth of Theory-of-Mind Reasoning in Sequential-Move Games. *Cognitive Science,* 36(3), 560-573. https://doi.org/10.1111/j.1551-6709.2012.01238.x

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2023). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *ICLR*. https://arxiv.org/abs/2205.10625

Vita

Haley Triem was born and raised near Houston, Texas. She holds a B.A. in English and Creative Writing and a B.A. in Art from the University of Iowa in Iowa City, Iowa. As an undergraduate student, she was awarded the Old Gold Scholarship, was on the Dean's List and the President's List, and graduated with University Honors.

Before entering the Graduate School at the University of Texas at Austin, Haley enjoyed various roles as a journalist, high school tutor, and library assistant. While at UT, she has worked as an archival project intern, a graduate assistant, a research assistant, and a teaching assistant. She has been awarded the Billie Grace Herring Scholarship and helped secure a position in the National Science Foundation's Innovation Corps (I-Corps) program, where she interviewed 120+ healthcare professionals as an Entrepreneurial Lead. She proudly serves as a Graduate Student Representative on the UT Informatics Undergraduate Studies Committee.

Haley is graduating with her MSIS from UT Austin's iSchool in May 2024 with an Endorsement of Specialization in Applied Machine Learning and Deep Learning, and a Graduate Portfolio in AI Ethics. She aspires to a life of learning and teaching.

LinkedIn: @haley-triem

Address: haleytriem@gmail.com

This dissertation was typed by the author.