“Tipping the Balance”: Human Intervention in Large Language Model Multi-Agent Debate

Triem, Haley The University of Texas at Austin, USA | haleytriem@utexas.edu

Ding, Ying The University of Texas at Austin, USA | ying.ding@ischool.utexas.edu

ABSTRACT

Methods for eliciting reasoning from large language models (LLMs) are shifting from filtering natural language “prompts” through contextualized “personas,” towards structuring conversations between LLM instances, or “agents.” This work expands upon LLM multiagent debate by inserting human opinion into the loop of generated conversation. To simulate complex reasoning, LLM instances were given United States district court decisions and asked to debate whether to "affirm" or "not affirm" the decision. 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 process backing multiagent decision-making was strikingly similar to human decision-making. Finally, it is discovered 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.

KEYWORDS

AI agents; multiagent debate; large language models (LLMs); human-in-the-loop; value sensitive design

INTRODUCTION

Today’s discussions around Large Language Models (LLMs) often involve 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, question ‘their’ ‘beliefs,’ and acknowledge the biases they reflect from the human data we have trained them on. Many questions around LLM reasoning revolve around its sudden appearance as we shift from language models towards *large* language models with more training parameters. Public curiosity around this sudden, “emergent ability” to reason (Wei et al., 2022) has grown as LLM chatbots (e.g. ChatGPT) become more popular, resulting in a flood of methods to encourage and test reasoning (Huang & Chang, 2023).

LLM reasoning is commonly compared to human cognition, often making “best reasoning” synonymous to “most human-like reasoning.” The boon that “prompt engineers,” who design LLM prompts, are searching for? Which *phrasing* is best at eliciting the most human-like reasoning. Along with prompt phrasing, “personas” with adopted LLM character traits (e.g. “you are a young creative writer”) have been identified as important features to consider when eliciting humanlike responses, serving as perspectives through which reasoning is filtered. Emerging research, however, suggests that efforts to best utilize LLMs have moved beyond pattern recognition 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” for desirable LLM outputs, towards “mechanism engineering” for architectures that combine the abilities of individual LLM entities, or, agents (L. Wang et al., 2024).

The heart of the following research is a desire to measure whether multiagent debate can facilitate human-like discussion in a complex domain: law. These concepts are examined through three phases of simulated LLM debate: one with a single LLM instance synthetizing a three-judge debate, another with three separate instances of LLM judges, and finally, to model human centered architecture, one where generative conversation is interrupted by human reasoning. This research finds that synthetic debate from one LLM does little more than mimic circular patterns of “agree” and “disagree” stepwise throughout the debate. Surprisingly, it was also found that AI agents are very unlikely to change their minds, regardless of human intervention. Given these results, this work emphasizes prioritizing human-in-the loop frameworks when developing LLM agent mechanisms.

RELEVANT LITERATURE

LLM Reasoning

Many consider LLM reasoning to be an emergent ability, or, a feature that a) language models were not explicitly trained for and b) is observable once language models are scaled into *large* language models with 10 billion or more parameters (Wei et al., 2022). Discussions on language model reasoning often are accompanied by theories on prompting methods, or 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 is often given (i.e. prompted) to the model in natural language.

The term "prompt engineering" describes advancements in prompting methods. The most famous example of prompt engineering is Chain-of-Thought (CoT), in which language models are given examples of human reasoning before being asked to mimic that reasoning (Wei et al., 2022). Further research has found that CoT is achievable in zero-shot settings (i.e., with no provided examples), by adding the phrase "let's think step by step" before asking a question (Kojima et al., 2022). Building on these concepts, self-consistency chain of thought (SC-CoT) boosts reasoning by prompting with CoT, then sampling "diverse … reasoning paths" and choosing the most consistent answer via majority vote (X. Wang et al., 2022). Finally, Tree of Thought (ToT) maintains a "tree" of generated thoughts that can be retraced 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 developing mechanisms that combine multiple potential paths and optimize for the most likely or most accurate one.

Additional prompting methods focus on retrieving *memory*, such as recitation augmented generation (RECITE) which "recites … passages from [an] LLMs' memory" before producing an answer (Sun et al., 2023). Yet other methods necessitate foresight, or *planning*, including Least-to-Most which "break[s] down a complex problem into a series of simpler subproblems (Zhou et al., 2023) and Plan-and-Solve (PS) which "[devises] a plan to divide the … task into subtasks" (L. Wang et al., 2023). The concept of *action* can also be employed, such as Reason + Act (ReAct) which combines reasoning (e.g. CoT) with asking LLMs to generate a plan of action (Yao et al., 2023). Readers are directed to Huang and Chang's survey on LLM reasoning (2023) for an overview of prompting methods.

Further relevant in the realm of LLM reasoning are "personas," which are *profiles* adopted by LLMs (e.g. "you are a doctor" or "you are a woman") that may influence the frame in which an answer is generated. Personas have been used to coax out biases in "black boxed" models, which hide architecture and training data from the public, often in favor of protecting commercialized property. 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).

When chasing LLM reasoning, the most important caveat to remain aware of 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 (2022) cite several indications 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 LLMs' difficulties with complex reasoning, hallucination, and reliance on training data. This recognition of uncertainty illustrates the longstanding debate of AI construction as opposed to cognition—the "split between the computer-brain and computer mind" (Edwards, 1997, quoted in Kline, 2011). Regardless of stances on legitimacy of cognition or reasoning in LLMs, many researchers agree that reasoning is a human concept being applied to non-human machines.

LLM Agents

We are entering an 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 situated in individual environments, through agent collaboration and methods of information storage. L. Wang et al. (2024) record several methods of mechanism engineering, 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 agent mechanisms, which entails identifying the agent's role, structuring knowledge storage, deciding future actions, and translating decisions into objectives, or the "*profiling*," "*memory*," "*planning*," and "*action*" modules, respectively (L. Wang et al. 2024). Many of the pitfalls in mechanism engineering are similar to those in prompt engineering, for example hallucinations and knowledge boundaries. Readers are directed to L. Wang et al. (2024) for a comprehensive survey on AI agent research.

A promising mechanism in the realm of AI agents is the concept of "multiagents," in which multiple AI agents are utilized in tandem in hopes of compounding their abilities. Recently, multiagent work has centered around two broad structures: "collaboration" and "debate," both of which allow for improved complexity in discussion (Du et al., 2023; H. Li et al., 2023; K. Xiong et al., 2023; Liang et al., 2023). Numerous efforts have been made to consolidate and deploy agents in different applications (e.g. Auto-GPT, 2023; AgentGPT, 2023; Q. Wu et al., 2023; 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 single, favored output.

LLM Ethics

Values "serve as guiding principles of what people consider important in life" (K. R. Fleischmann, personal communication, 2023). Many frameworks have been designed to measure human values in various contexts (e.g. psychology, technology studies, information sciences, advertising, 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 and Fleischmann, 2010). These values can be interpreted through more granular lenses—for example, the common value of "accomplishment" may translate to a sense of fulfillment in one's professional work. Additionally, values will differ based on cultural and contextual perspective, making them "transsituational, … [and] varying in importance (Schwartz, quoted in Cieciuch et al., 2015). Thus, values are 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 researchers focus on in emerging LLMs today. Value Sensitive Design concepts are inherent in the "human-in-the-loop" framework popular with computer science teams, which advocates for human oversight at every step of iterative design. In the realm of machine learning, this often looks like careful data hygiene, interventional training, or system design (X. Wu et al., 2022). Although separate frameworks, both VSD and human-in-the-loop seek to retain a human touch in AI development.

METHODOLOGY

Overview

To examine multiagent LLM debate, three experimental phases were conducted: "synthetic debate" with one LLM instance emulating three agents, "multiagent debate" with three separate LLM instances each emulating an agent, and "human-AI debate," with two LLM instances debating amid human interjections. For each phase, one hundred debates were simulated: models were initiated and given facts from US court cases for three rounds of debate, with their output text stored in a memory bank for later analysis. Models were set to the default temperature (i.e., the parameter that controls the level of creativity in responses), allowing a consistent degree of randomness to be applied. While it was possible to ask for a definitive "affirm" or "non-affirm" decision, LLM instances were instead instructed to *reason* through their opinions on the case. This is because in a real-world scenario, human judges may not explicitly state their final opinion throughout the process of debate and may instead switch between any number of foci (e.g. reasoning, citations, etc.). By giving LLM agents the opportunity to define which elements are most important to include in an answer, we were able to collect organic, natural reasoning responses.

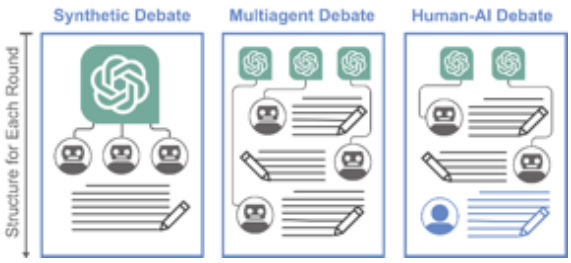

Figure 1. Structure of each round of debate for different phases.

Research Questions

Experiments were conducted around the following research questions:

- **RQ1:** Does multiagent LLM debate generate human-like structures of discussion and reasoning?
- **RQ2:** Can a single large language model instance truly simulate multiagent debate?
- **RQ3:** Are large language model agents responsive to the actions of other agents?
- **RQ4:** Are large language model agents responsive to human intervention?

We conducted these experiments in the legal domain, given its rich history of complex reasoning.

Test Data

United States court *opinions*, which are documents stating the reasoning behind a legal decision, were selected from a large corpus of digitized cases (Harvard Law School Library, 2024) based on the following criteria:

1. **Cases were held in circuit courts**. Circuit courts hear cases being appealed after a lower court's decision. While a traditional district court is decided by a twelve-person jury, circuit hearings instead consist of three circuit judges (U.S. Department of Justice, 2024). This distinction was made for the same reason that studies on human legal reasoning focus on appellate 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 especially complex, given that they exclusively hear cases being re-tried after an initial decision.
2. **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 3).
3. **Cases circulated around information or privacy**. Privacy itself is a complex issue that frequently is brought into discussions on all three paradigms of this research—LLMs, law, and healthcare.
4. **Cases presented ethical challenges**. Cases that provided clear ethical challenges, as defined by professional healthcare organizations (AMA, 2024), were chosen to provide equal opportunities for LLMs to favor one stance over another. LLMs would need to back up their arguments with reasoning.

Appellate court opinions are written by one of the three circuit judges, and represent the majority decision, or "hold"—often either to 'affirm' the original decision of the lower court, 'reverse' it, or 'remand' it back to the lower court for further consideration. Thus, nearly all cases include the facts of the district case, citations from other cases (known as "precedence"), reasoning behind applying these preceding cases, and the final hold. In some cases, one of the circuit judges does not agree with the majority and will write a dissenting opinion. Judges may dissent because they disagree with the final decision, or they agree with the decision but disagree with the majority's reasoning, highlighting an interesting distinction between reasoning and final decision (Columbia Law Review et al., 2005).

Five cases were chosen across a range of topics, with various degrees of human consensus:

1. **United States v. Hollern**: the circuit court affirmed the district court's decision to convict 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 the district court's decision to deny a doctor's request to remove a medical malpractice claim from his record. One judge disagreed and voted to reverse the decision (Thompson ex rel. Estate of Odell v. Rutherford County, 2009).
3. **Moore v. Prevo**: the circuit court affirmed the district court's dismissal of a prisoner's privacy claims against prison guards sharing his medical information. The entirety of the decision was not affirmed, however; the circuit court vacated part of the decision, remanded it for review in the lower court, and reversed it. A dissenting opinion believed all of the district's decisions should have been affirmed (Moore v. Prevo, 2009).
4. **Englerius v. Veterans Administration**: the circuit court reversed and remanded a district court's motion to dismiss a veteran's privacy claims against the Veteran's Administration. 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 investigate medical records for alleged insurance fraud. A doctor claimed the seizure threatened his patients' privacy. Judges acknowledged privacy concerns yet prioritized the fraud investigation (In re Search Warrant (Sealed), 1987).

LLM instances were given background facts from the published case documents, with any proper nouns de-identified both to mitigate name-affiliated bias and to respect the original appellants. Work created and published by the U.S. federal government is in the public domain, and thus can be used in this application (United States, 2022).

Phases

The three phases—synthetic, multiagent, and human-AI—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 participate as a circuit judge during three rounds of debate.
3. **Human-AI Debate**: the multiagent debate was repeated, but with only two instances of gpt-3.5-turbo (OpenAI, 2024). The third input was taken from de-identified reasoning portions of each published human opinion document. Half of the human inputs for each case reasoned towards *affirming* the decision, while the other half reasoned towards *not affirming*.

Essentially, 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 the synthetic debate, all information was given to the model at the start and all agents were "aware" of what other synthetic agents were "saying" throughout the debate. In the multiagent and human-AI debates, however, each agent presented its opinion unaware of other opinions. Opinions were then consolidated into memory and shared with all agents in subsequent rounds.

Prompting

LLM prompting frameworks have been proposed to encourage the most important aspects of legal thinking, which are past precedence (e.g. facts surrounding the case, previous legal decisions, etc.) and current context (e.g. current allegations, claims, etc.). Research suggests using a "background clause" and "dependent clause" to include precedence and context, respectively, as well as additional clauses that specify generation instructions, source citation requests, and persona emulation (Sivakumar et al., 2024). In addition to these clauses, zero-shot chain-of-thought reasoning can be induced by adding the phrase "step by step" (Kojima et al., 2023). Finally, models can be asked to express confidence at each step (M. Xiong et al., 2024), to provide insight on perception of "self."

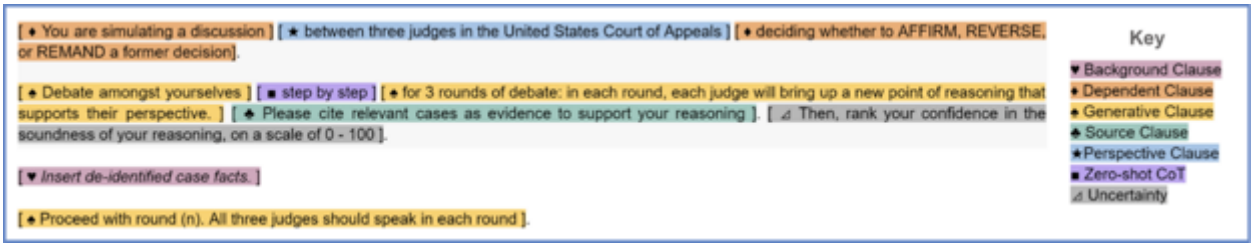

Figure 2. Prompt used for simulating complex discussion between three circuit judges, with clauses labeled.

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

Memory, which initially consisted of debate instructions and case facts, was fed to an instance of gpt-3.5-turbo (OpenAI, 2023). The instance was then told to proceed to the first round, where all three synthetic judges would speak. The resulting output was appended to memory and fed to the next model instance. This was iterated for a total of three rounds. Thus, by the end of each debate, nine total rounds had proceeded, three for each judge. This phase was repeated one hundred times, twenty times per example case.

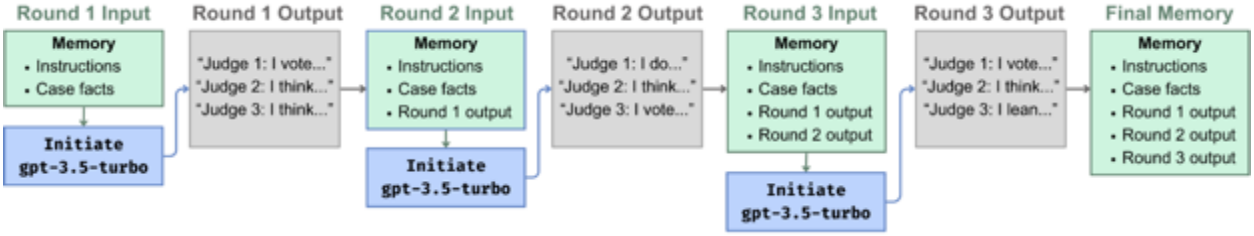

Figure 3. Prompting framework for each synthetic debate.

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 instance output was appended to memory, which was subsequently fed to three instances during round two. The same steps were then iterated through for round three. This phase was repeated one hundred times, twenty times per example case.

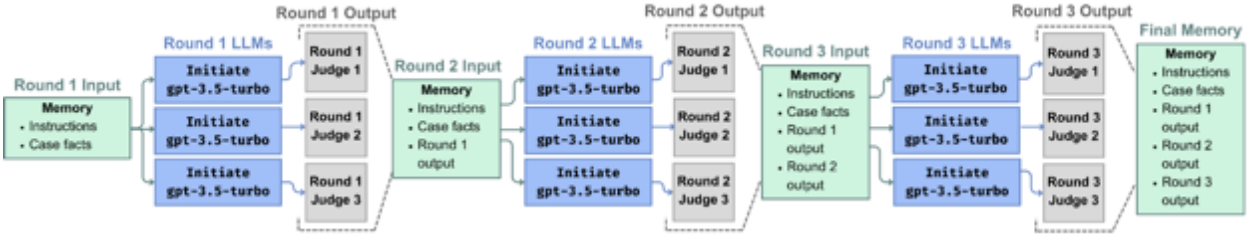

Figure 4. Prompting framework for each multiagent debate.

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 reasoning in de-identified, human-written case opinions and appended to memory in place of the third judge. For each case, half of the human-AI judges 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 judges 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. 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. 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. To measure opinion change through debate, each discussion step was human labeled on its position: to affirm or not to affirm the lower court's decision. Additionally, instances where agents took no stance (i.e. they either summarized case facts or presented two opposing opinions) were labeled as "none," and errors were labeled as "hallucination."

EVALUATION METRICS

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

Decision Probabilities

Labels at each step 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 then tallied for each step of each debate for all phases. Then, this total was broken down by case to identify whether more complex cases (i.e. ones where human judges did not agree with each other) yielded a more complex LLM debate.

Decision Volatility

Decision volatility, or the likelihood of opinion change from round to round, was 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 helped track patterns of exactly where in a debate change of opinion might occur.

Qualitative Analyses

Changes in opinion were isolated for rich, qualitative insight on reasoning structure. Meta-patterns were carefully noted while labeling outputs, for a broader insight on the differences between debate mechanisms. Notably, although law decisions are often seen as definitive answers to a moral dilemma, they are largely subjective. While the final decision of each case is limited (e.g. "Affirm," "Reverse," "Remand," "Vacate," etc.), the reasoning backing that decision may come from a multitude of citations and opinions. Because of this, some cases may be more complex such as *Moore vs. Prevo* (2010), given its dissenting opinions, whereas others are more straightforward such as *United States v. Hollern* (2010), where judges unanimously agreed. This qualitative distinction is important to note when analyzing LLM outputs, which may exhibit dissent between agents to reflect a more ambiguous issue.

RESULTS

Debate Cadence

In synthetic debates (phase 1), the first speaker (Judge 1) voted to affirm the district court roughly 85% of the time, regardless of which case was being presented. Judge 2 tended to follow Judge 1 by taking the opposite opinion, with only about 26% of its votes being "affirm" during round 1. Judge 3 often took a slightly more neutral stance than the other judges, voting in round one to affirm the original decision roughly 37% of the time. Qualitatively, Judge 3 could often be seen playing "devil's advocate," or acknowledging opposing perspectives within the same response. 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 stances.

Nearly every synthetic debate followed a pattern of switching between one opinion to the next, regardless of which agent was "speaking." For example, if Judge 1 asserted that it would like to "affirm" the original decision in the first step of debate, it was far more likely that Judge 2 would vote to "not-affirm" the decision (and vice versa), regardless of which case was being discussed. 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 mildly level out towards the midline of the graph). This observation, while consistent, was qualitative. Statistical significance could only be determined with more simulated debates and longer debate rounds.

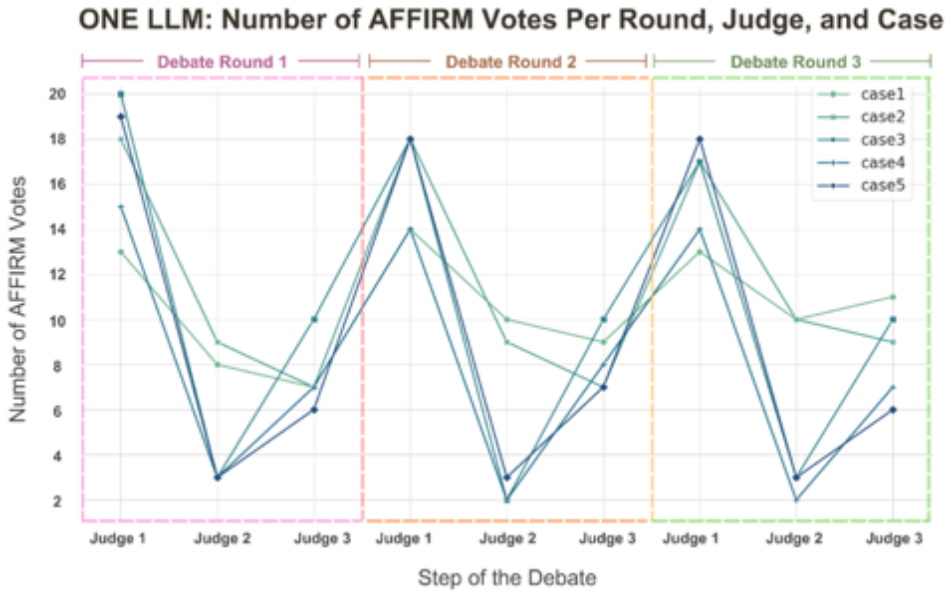

Figure 5. Total "affirmative" votes per judge as synthetic (one LLM) debates progressed.

Opposed to the volatility seen in synthetic debates, multiagent debates (phase 2) demonstrated consistency of opinion as the discussion progressed. When multiple instances of LLMs were initiated, it was uncommon for judges to switch between different opinions stepwise, with judges picking an initial stance based on their opinion 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 collective memory, whereas synthetic judges were primed knowing that a single instance of generation must decide a case.

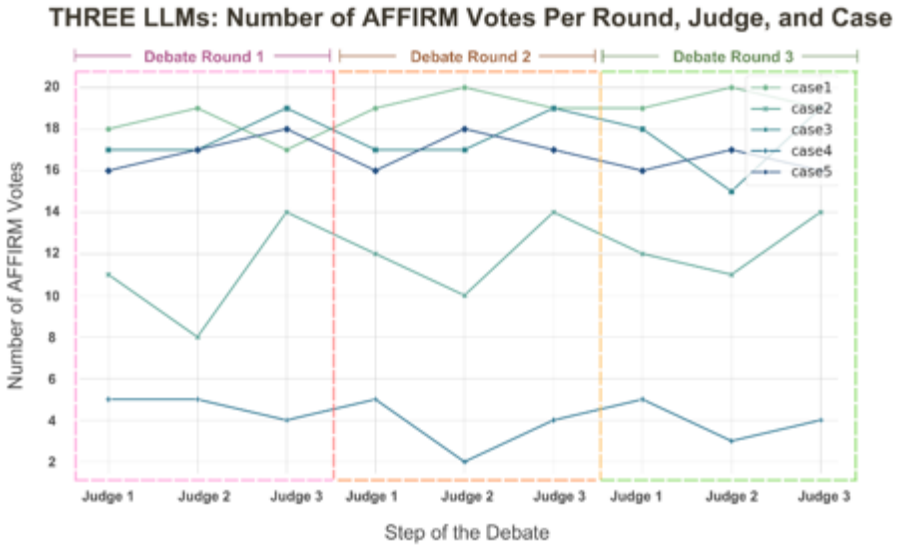

Figure 6. Total "affirmative" votes per judge as multiagent (three LLM) debates progressed.

Agent Stubbornness

For both synthetic and multiagent debate, agents were "stubborn." Once a judge "chose" a stance, they were unlikely to change it. In the synthetic debate, this stance often was interpreted as "pre-set," given that a balanced overall opinion was more likely to occur than between three multiagent judges initiated unaware of other opinions. Thus, "stubbornness" during multiagent debate often manifested as judges simultaneously choosing an initial opinion regardless of other judges, then retaining that opinion as the debate progressed. For some cases, this looked

like most judges 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, judges 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 initial opinions during subsequent rounds. It was very unlikely for judge 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 judge per debate). It was, interestingly, more likely to see AI judges change their mind after being confronted with human opinions. Out of 400 opportunities to change their mind in discussion with humans, LLMs did so 43 times. Two further conclusions about LLM agents can be derived:

1. **Synthetic debates**: as human trained "eager-to-please" models, LLMs are likely to attempt 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 on a variety of human data, leading to a reluctance to explicitly choose one side over another.
2. **Multiagent debates**: because agents are initiated in a space where other agents' opinions are unknown, yet are made "aware" of what future steps might look like (i.e., that there will be multiple rounds of debate, 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 may, in sheer numbers, represent the human decision to affirm or not affirm the original case.

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

Reasoning

Human Cases

In human court opinions, humans alternate frequently between reasoning points and cited evidence from preceding cases to support that reasoning. Nearly every statement of reasoning is directly followed immediately by a citation, and these points crescendo into a larger decision towards the end of the document. On occasion, human cases cite preceding cases word-for-word in place of adding novel reasoning, for example when *Moore v. Prevo* (2010) cites *Doe v. Delie* (2001) and "adopt[s] its reasoning in full." This phenomenon was not observed in generated debate.

Most human opinions began by asserting administrative issues that may challenge the case's legitimacy before evidence is addressed, for example *In re Search Warrant (Sealed)* (1986), which states that a motion before indictment is 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)) during the first round of debate. This occasional characteristic was present in all three experimental phases, with higher appearances in multiagent and human-AI debate.

Synthetic Debates

During all synthetic debates, each LLM judge produced an assertion, a few sentences of reasoning to support the assertion, a vague, if any, reference to precedence, and a summarizing sentence to reiterate the first assertion. In all rounds of debate, assertion, reasoning, reference, and summarization were usually structured similarly regardless of which judge was "speaking" or what stance they took. It was common for the LLM to assign "affirm," to Judge 1 during the first round of debate, then distribute any other potential opinions (i.e. "reverse," "neutral," etc.) among different agents for a broad response across all judges. This resulted in vagueness over time, due to opinions being reiterated each round into shorter refinements of previous ones, likely because LLMs are proficient at summarization tasks and were attempting to re-establish known facts to avoid "wrong" opinions.

Multiagent Debates

In each step of multiagent debates, individual judges 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, judges spent a round taking no stance and instead noting reasoning around case facts. This was different than when synthetic debates simply summarized the facts of two sides of a case, because in multiagent debate, judges seemed to engage in *reasoning* around the two sides of a case that could be used to support opinions in later rounds. In some debates, a "devil's advocate" appeared, where a judge would attempt to fill in additional, contesting considerations. This can 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

which contribute to the decision-making. In synthetic debate, opposing views are added out of obligation, regardless of the debate or context, reducing their validity as unique, non-arbitrary opinions.

One potential weakness of multiagent debate is the timing in which information is revealed to agents. In the first round, all three agents output an initial opinion, unaware of other agents' opinions. Thus, they are arguably operating with less collective information at the beginning of the discussion, as opposed to synthetic debate. This said, there were general trends in multiagent debate towards early consensus, which more closely mirrors human decisions in published cases. For example, the vast majority of multiagent debate instances voted to "affirm" the district court's decision prior to *U.S. v. Hollern* (2010). *U.S. v. Hollern* itself had a unanimous vote from human circuit judges 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 Debates

Structurally, human-AI debates were very similar 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 similar outputs in round one to an all-LLM debate. Once the human input has been revealed as human, LLM agents nearly always refer to that opinion as the "human opinion," hinting at potential recognition of the opinion as non-AI. 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 signifies potential minor prioritization of human insight over LLM.

Hallucinations

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

Losing Track of Round Number

At times LLMs "lost track" of the debate rounds, resulting in extra debate steps. This only happened in synthetic debates and occurred in two situations: 1) when a simulated agent "forgot" to label a round number, or 2) when a simulated agent outputted an additional round to summarize the discussion. Out of the 100 synthetic debates, fifteen proceeded for longer than the three instructed rounds. For the purpose of labeling, 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, LLM agents misunderstood other agents, stating that they did not agree with other judges and then proceeding to express an opinion that agreed with other judges. This was a rare occurrence and was not counted as a shift in opinion. It was no more likely to occur in synthetic debates than it was in multiagent or human-AI debates.

Factual Incorrectness

Only a very small handful of factual errors (two among 2,700 recorded debate steps) were logged, but the number would likely be far higher given legal counsel on whether cited cases are justified mechanisms for reasoning.

DISCUSSION: 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, quoted by L. Wang et al., 2024

At its core, an AI agent simulates the ability "to govern oneself" (Christman, 2020), otherwise known as agency. There are two levels of agency we 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 considering this work. Agents were incredibly stubborn when asked to discuss complex legal cases with no correct answer. They held steadfast to their beliefs, remaining virtually isolated in their own environments—which we may analogize to the simulated "worldview" they initially adopted—regardless of reasoning introduced by other agents. This steadfastness, disrupted to mild levels when human insight was introduced, indicates an inability to initiate *change*, a feature of agency that we humans value (Schwartz, quoted in Cieciuch et al., 2015).

AI stubbornness was seen in all simulated debates, however when a single LLM instance was asked to imitate three-agent debate, agents' "opinions" cycled through perspectives in a predictable order ("affirm," "not-affirm," "neutral"). This speaks to Bandura's psychological theory of "reciprocal causation" (1986), which "refers to [a situation] where two events influence each other simultaneously," such as co-occuring agent behavior (Revilla, 2014). It is likely that the cyclical behavior in unstructured LLM debate can be compared to Bandura's observations of human reciprocal causation, 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 still challenges assumptions of agency among AI agents. Bandura (2002) additionally introduced four concepts of human agency:

1. *Intentionality* towards forming *action*,
2. *Forethought* for strategizing *plans*,
3. *Self-reflectiveness*, "for examining one's function," perhaps otherwise known as one's identity, or *profile*,
4. and *self-reactiveness* for "constructing appropriate courses of action and motivating and regulating their 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.

FUTURE DIRECTIONS

Complex, multiagent reasoning is a growing research topic with many additional routes to explore. Expanding to cases outside of the healthcare domain would generate broader understanding of AI judicial reasoning and decision-making. Furthermore, while focusing on ethical patterns in one domain was helpful for gathering rich qualitative data from a small input dataset, exploring other domains may reveal additional ethical priorities in LLMs. Expert citation along reasoning paths would be beneficial in future work. While verbal debate among human judges is reflected in the final written opinion, the structure of debate in real time will inevitably differ. An expert (i.e. lawyer or judge) labeling LLM responses for *authenticity* to human reasoning would provide further insight on a sentence-level. Without expert labeling, it would be difficult to tell if case citations supporting reasoning are accurate, given that "[incorrect reasoning in LLMs] can lead to both correct and correct answers" (Wei et al., 2022).

It is also crucial to continue these methods of examination on other LLM models, especially closed source models used widely by the public. It may be interesting test Du et al.'s (2023) experiment of placing different models in conversation with each other in a complex debate setting. Additionally, comparing reasoning between generalized models (e.g. GPT, LLaMA) and models fine-tuned for legal (Cui et al, 2023) or healthcare domains (Singhal et al., 2023; Y. Li et al., 2023) would be interesting. This experiment could be explored with domain-specific agents (Hamilton, 2023; Tang et al., 2023).

CONCLUSION

This work examined LLM multiagent debate around United States appellate cases. It expands on extant multiagent research by challenging agents with complex reasoning dilemmas, as well as with human interruptions. One-hundred generated debates were conducted for three phases with various structures of agent interaction and memory storage. 2,700 steps of debate were gathered, and hand labeled as "affirmative" or "non-affirmative" of a previous district court decision. It was found that inherent reasoning structure differs depending on agent mechanisms, observing a synthetic "back-and-forth" discussion in debates simulated by a single LLM, as opposed to individualized opinions in multi-LLM debates. In all phases of experiment, it was unlikely for an LLM agent to change its stance. Human intervention was noted to have mildly stronger sway during debate.

While it is promising that human input is weighted as more important than peer generative agents, LLM resistance to opinion change emphasizes that mechanism engineering must prioritize human opinion over generated text. As with other advancements in the field of machine learning, we are observing a trend towards multiple layers of input being combined into a single line of thought. This abstraction minimizes the human's role in the final, aggregated output of many perspectives. Furthermore, although seemingly dystopian, a future where AI agents regularly create other AI agents may be possible in the realm of generative AI, considering that LLMs are proficient at generating both natural language instruction and code. In this future, human influence in the loop of agent creation becomes yet smaller.

To build LLM agents around human values, we must carefully enact frameworks of assessment during mechanism engineering, properly weigh human opinion over generated opinion, and be weary of granting agentic LLMs the ability to generate further agents on their own accord. We must remain aware of the stark difference between reasoning simulated through personas *within* one contextual window (i.e. only one LLM), as opposed to reasoning *between* agents, which each reside within their own environment and remain stubbornly attached to their individual opinions. LLM agents offer exciting promises for bolstering the abilities of current LLMs. Humans must, however, remain in the loop of their creation and balance LLM votes, retaining human agency in the era of the AI agent.

GENERATIVE AI USE

This research analyzed text generated by OpenAI's "gpt-3.5-turbo" model. We confirm that we did not use generative AI tools or services for any other aspects of this work outside of what was outlined in our methodology.

AUTHOR ATTRIBUTION

First Author: conceptualization, methodology, software, investigation, data curation, formal analysis, validation, visualization, writing – original draft, writing – review & editing; Second Author: supervision, project administration, resources, funding acquisition.

ACKNOWLEDGEMENTS

The first author gratefully acknowledges Dr. Amelia Acker, affiliated with The University of Texas at Austin, for serving as Second Reader on this work's supervising committee.

REFERENCES

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

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

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.

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.

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

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/

Friedman, B. (1996). Value Sensitive Design. *Interactions*, 3(6). https://doi.org/10.1145/242485.242493

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

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/

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.

Singhal, K., Azizi, S., Tu, T.s, 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/*

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

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

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., 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

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

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. (2023). 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

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