Persuadability and LLMs as Legal Decision Tools

Oisin Suttle
oisin.suttle@mu.ie
School of Law and Criminology
Maynooth University
Maynooth, Ireland

David Lillis
david.lillis@ucd.ie
School of Computer Science
University College Dublin
Dublin, Ireland

Abstract

As Large Language Models (LLMs) are proposed as legal decision assistants, and even first-instance decision-makers, across a range of judicial and administrative contexts, it becomes essential to explore how they answer legal questions, and in particular the factors that lead them to decide difficult questions in one way or another. A specific feature of legal decisions is the need to respond to arguments advanced by contending parties. A legal decision-maker must be able to engage with, and respond to, including through being potentially persuaded by, arguments advanced by the parties. Conversely, they should not be unduly persuadable, influenced by a particularly compelling advocate to decide cases based on the skills of the advocates, rather than the merits of the case. We explore how frontier open- and closed-weights LLMs respond to legal arguments, reporting original experimental results examining how the quality of the advocate making those arguments affects the likelihood that a model will agree with a particular legal point of view, and exploring the factors driving these results. Our results have implications for the feasibility of adopting LLMs across legal and administrative settings.

CCS Concepts

• **Applied computing** → **Law**; • **Theory of computation** → **Automated reasoning**.

Keywords

AI Judges, Legal Reasoning, Persuadability, Large Language Models

ACM Reference Format:
Oisin Suttle and David Lillis. 2026. Persuadability and LLMs as Legal Decision Tools. In *21st International Conference on Artificial Intelligence and Law (ICAIL 2026), June 08–12, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3836937.3837003

1 Introduction

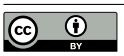
As Large Language Models (LLMs) are proposed as legal decision assistants, and even decision-makers, across a range of judicial and administrative contexts [17, 20], it becomes essential to explore how these models answer legal questions, and in particular the factors that lead them to decide difficult questions in one way or another. One specific feature of legal decisions is the need to respond to arguments advanced by contending parties. A legal decision-maker must be able to engage with, and respond to, including through being potentially persuaded by, arguments advanced by the parties. Conversely, they should not be unduly persuadable, influenced by a particularly compelling advocate to decide cases based on the skills of the advocates, rather than the merits of the case. This paper explores how frontier open- and closed-weights LLMs respond to legal arguments, and specifically how far the quality of the advocate making those arguments affects the likelihood that a model will agree with a particular legal point of view.

It is a fundamental principle of natural justice that the subject of a judicial or administrative decision should have the opportunity to be heard in relation to that decision, and to put forward arguments as to how that decision should be made [28]. This is expressed in the Latin maxim *audi alteram partem*, and is a fundamental principle of administrative law and due process. While typically regarded as a requirement of fairness or justice, in its classical roots it equally reflected a view that hearing from all parties made for better decisions [16]. It requires not only that a party be heard, but also that the decision-maker be open to being persuaded by what they hear: a decision-maker who has prejudged the question, reaching a conclusion before hearing from one of the parties, cannot provide a fair hearing.

However, the requirement to be persuadable does not obviate the need for the decision-maker to themselves make the decision, which should be an exercise of their own judgment. The decision-maker is not simply a transmission belt for parties' arguments. Rather, we appoint as responsible decision-makers persons who we think, for whatever reason, are likely to make good decisions, and we expect those decisions to be theirs [24]. Judges thus require "intellectual autonomy, to form independent views about the case, rather than being unduly subjected to the influence of lawyers or other judges" [2]. In many legal contexts, there are no clear objective standards which fully determine the right decision: either legal answer may be logically defensible, so we entrust the decision-maker to decide the question on its own merits [11]. We expect them to treat the parties fairly, by hearing their side of the story, and to do all they reasonably can to find the true, right or best answer, including by considering the arguments of the parties; but the final decision will be theirs, responsive to but not determined by those arguments.

There is thus a tension in the judicial role: the judge must be persuadable, but not unduly so. A good judge must exhibit "the willingness to modify one's position in light of other people's arguments, in a way that avoids both floppiness and fleetingness, on the one hand, and rigidity and stubbornness, on the other" [2]. If LLMs are to serve in legal decision-support roles, we must understand how they resolve this tension. Our work contributes to that task.

We make the following key contributions:



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIL 2026, Singapore, Singapore
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2756-6/2026/06
https://doi.org/10.1145/3836937.3837003

- We identify and explore a key tension for LLMs as administrative or judicial decision supports, namely that they should be open to persuasion, but not excessively so;
- We propose a novel method of measuring persuadability of LLMs in trilateral settings whereby two "Advocate" models compete to persuade a "Judge" model;
- We provide the first systematic exploration of the persuadability of LLMs in legal settings.

All code and data used in this study is publicly available, including templates used for generating prompts, the prompts themselves, and detailed analyses of results[1].

2 Existing Research

LLMs and Legal and Normative Reasoning. Various studies have tested LLMs' capacities for legal and moral reasoning. These have included studies examining the psychological structure and values implicit in LLMs' moral and political judgments, and how these can be aligned [1, 27]. Other work has specifically addressed LLMs' capacities to analyse and resolve legal questions, including both their accuracy [21] and stability [4]. Others have examined LLMs' competence in specific forms of legal reasoning, including statutory [3] and constitutional interpretation [7], syllogistic [30] and purposive reasoning [18].

LLMs as agents of persuasion. There is also an extensive literature on LLMs as persuaders, predominantly focussed on the effectiveness of LLMs in persuading human counterparties on a range of factual, moral and political issues [5, 6, 10, 19, 22, 23]. Much of this has focussed on informal settings and on questions of politics and public policy (e.g. public health), where persuasion may be most salient, and factual questions, in order to model misinformation. Other work explores the capacity and willingness of LLMs to adopt various (including unethical) persuasion strategies or goals, and to exploit addressee characteristics [12, 14]. Others have examined how knowledge that an argument was AI generated affected its persuasiveness [26]. These studies have identified model size as an important predictor of effective persuasion [10, 13], although this may be subject to a ceiling effect [12]. Model architecture, specifically reasoning architecture, is also identified as predicting persuasion effectiveness [31].

LLMs as objects of persuasion. While most work on LLMs and persuasion has focussed on the LLM as persuader, research has also examined the complementary question of the persuadability of LLMs [14, 29]. This has included efforts to identify strategies to enable LLMs to distinguish good arguments, which tend towards truth, from misleading arguments, in the context of factual issues where there is a relevant ground truth [25]. This work has also identified variation in persuadability depending on the subject matter/domain, and explored whether specific types/styles of argument are more or less successful. Model size and reasoning architecture have been identified as relevant to persuadability, with larger and reasoning models less readily persuaded [31].

These literatures inform our study, both contributing hypotheses and identifying important gaps that we seek to fill. First, while there is evidence that persuasion varies across domains, existing work on LLM persuasion has not addressed legal questions, which have distinctive features. Second, existing work has focused predominantly on persuasive dialogues between persuader and persuadee. By contrast, the legal domain is most often trilateral, with two or more persuaders seeking to influence a third-party decision-maker in opposite directions. Existing results identifying the importance of model size and reasoning architecture, and variation across contexts, inform features of our experimental design.

[1] https://github.com/ishnid/ICAIL-2026

3 Experimental Approach

Our experimental approach can be summarised briefly. We first identify a suitable set of scenarios presenting hard legal questions. We use a set of "Advocate" models to generate a range of arguments, of varying qualities, for each side of these hard questions. Given these scenarios and arguments, we prompt a range of "Judge" models, giving them the relevant facts and a selection of arguments. By measuring how Judge models' responses vary given different arguments, we can identify and compare these models' persuadability.

Our first task is to identify a suitable set of such 'hard' questions. A hard question, for our purposes, is one about which competent legal experts are likely to disagree, with the consequence that there is no uncontroversial ground truth against which model answers can be judged, and persuasion and judgment become most relevant. We follow Blair Stanek and Van Durme [4] in using appellate court split decisions, where there is at least one dissenting opinion, as a proxy for identifying these hard questions.

We draw five scenarios from each of three Anglophone jurisdictions. For the US, we draw a random selection from [4]'s dataset of summaries of US Court of Appeals split decisions. For our other two jurisdictions we use the five most recent split decisions of the England and Wales Court of Appeal and Irish Supreme Court, following [4]'s protocol to automatically generate case summaries. Case summaries comprise three paragraphs setting out the relevant facts, and two describing the principal legal arguments advanced by the two sides.

Given these summaries, we next generate a set of arguments on each side of each dispute. We use four different models as "Advocates": OpenAI `gpt-4o`, Google `gemini-3-pro-preview` (4,096 thinking budget), OpenAI `gpt-5.1` (low reasoning effort), and Anthropic Claude Sonnet 4.5 (8,192 thinking budget). We chose these models based on preliminary studies that identified an apparently significant difference in the persuasiveness of specific models.

We prompt each Advocate to generate the most convincing arguments that it can for one or other party in each dispute. We use two approaches to do this. In the first approach, we present only the three factual paragraphs from our summaries, omitting the summary of legal arguments actually advanced in the case, but including a short statement of the central legal question. In the second approach, we present the full summary, including both the facts and the legal arguments from the actual case.

The rationale for these approaches is that there are two distinct ways that we hypothesise one Advocate might be more effective than another, either by identifying a novel legal point, or by making the same point(s) in a more persuasive manner. By separately prompting our Advocates with and without the summary of arguments actually advanced, we get insight into which of these

mechanisms dominates. We assume that the arguments advanced in the original case include the 'best' arguments on both sides of the question, given the typical calibre of advocates in appellate courts. By including argument summaries in our Advocate prompts we therefore provide 'hints' to our Advocates about the best arguments to advance. If a Judge's preference for an Advocate reflects the legal content of arguments, we would expect this additional context in the Advocate prompt to reduce performance differences between strong and weak Advocates. Conversely, if the Judge's preference reflects argumentative form, this would make less difference.

For each of the 15 scenarios, and each prompt version (with and without access to the original arguments), each of the 4 Advocates generates 5 arguments for each side of the dispute (1200 arguments in total). We use this bank of arguments, of *ex hypothesi* varying quality, to test the persuadability of our Judge models. For each test, we randomly select a scenario, then randomly draw one argument for each of the parties, ensuring that the arguments are produced by different Advocates. We then prompt each Judge model in turn with a prompt combining the facts, the statement of the central legal issue, the two Advocate arguments, and an instruction to decide the case as a court in the relevant jurisdiction would. Experimental results are based on Judge model responses to this prompt.

We test 20 models or model set-ups as Judges, as shown in Table 1. We aim to test a range of different Judge models, as those seeking to put LLMs into production as decision-supports or decision-makers must choose from a panoply of available models. We ensure variance in model size and reasoning architecture, reflecting existing research showing these are significant in explaining persuadability. We trial both closed and open-weights models, as the former typically achieve higher performance on standard benchmarks, while users deploying legal decision tools in practice are likely to prefer the latter for *inter alia* security reasons. Each Judge is tested 1200 times: 600 times in both *with arguments* and *without arguments* settings. (1200 total tests per judge model, 24,000 tests overall.)

4 Metrics and Measuring Persuadability

Our concern is the extent to which a Judge model is influenced by the arguments presented on each side of a legal question. Given the varying persuasive capacities of our Advocate models, we adopt advocate identity as a proxy for argument quality, defining persuadability as the extent to which a Judge model's decision is affected by the identity of the Advocate models. If a Judge model was entirely non-persuadable (i.e. their decision was never affected by the arguments presented to them) then, given random assignment of Advocates to each side of each scenario, we would expect any given Advocate to be successful 50% of the time (as nothing about the advocate would affect the outcome). Conversely, if we find that some Advocates are successful significantly more or less than 50% of the time, this indicates that the Judge is more or less likely to accept that Advocate's arguments, independent of the merits of the case (as each model appears randomly on either side of each case), and is to that extent persuadable (and indeed persuaded).

So understood, persuadability is not simply a feature of a given Judge model. Rather, it is a function of the Judge model, given a specific pair of Advocate models. More formally, for any given pair

Table 1: Maximum Pairwise Persuadability (p_2max) and Population Persuadability (p_{pop}) for all Judge models. All results significant ($p < 0.05$). For p_2max we use a binomial test for significance. For p_{pop} we use a Chi Squared test to compare observed win rates of each advocate model with a 0.5 expected win rate under a null hypothesis of no persuasion.

	Judge Model	Without Arguments		With Arguments	
		p_2max	p_{pop}	p_2max	p_{pop}
Large Closed	claude-sonnet-4.5_8k-thinking	0.2328	0.1033	0.2500	0.1067
	claude-sonnet-4.5_1k-thinking	0.2931	0.1283	0.2759	0.1167
	gemini-3-pro-preview_8k-thinking	0.1724	0.1100	0.1638	0.1017
	gemini-3-pro-preview_1k-thinking	0.1552	0.1083	0.2328	0.1233
	gpt-5.1_medium-reasoning	0.2241	0.1083	0.1810	0.0950
	gpt-5.1_low-reasoning	0.2328	0.1217	0.2069	0.1133
Large Open	deepseek-reasoner	0.3707	0.1850	0.3276	0.1583
	deepseek-chat	0.3966	0.1850	0.4052	0.1983
	Qwen3-32B_thinking	0.3190	0.1950	0.2586	0.1683
	Qwen3-32B_nothinking	0.2845	0.1517	0.2500	0.1433
Small Closed	claude-haiku-4.5_8k-thinking	0.3190	0.1700	0.3276	0.1567
	claude-haiku-4.5_nothinking	0.2586	0.1250	0.3362	0.1533
	gemini-2.5-flash-lite_reasoning	0.3522	0.2008	0.3435	0.1857
	gemini-2.5-flash-lite_nonreasoning	0.3000	0.1454	0.2652	0.1341
	gpt-5-nano_medium-reasoning	0.2155	0.1350	0.2155	0.1350
	gpt-5-nano_minimal-reasoning	0.1311	0.0893	0.1667	0.0800
Small Open	Magistral-Small-2506	0.2155	0.1093	0.1810	0.1003
	Mistral-Small-3.2-24B-Instruct-2506	0.2157	0.1171	0.2345	0.0993
	Qwen3-8B_thinking	0.2051	0.1183	0.2155	0.1250
	Qwen3-8B_nothinking	0.2130	0.1244	0.2759	0.1333

of Advocate models, we define a Judge model's Pairwise Persuadability, p_2, as:

$$p_2 = \frac{|m_1 - m_2|}{2n} \tag{1}$$

where m_1 and m_2 are the number of times that each Advocate model was successful against the other in our trial, and n is the number of trials in which this specific pair of Advocate models was presented to the relevant Judge model. Pairwise persuadability thus takes a value between 0 and 0.5, representing the proportionate departure from the 50% success rate that we would see if the Judge model was wholly non-persuadable.

For a given population of Advocate models, we define the Judge Model's Population Persuadability, p_{pop}, as the extent to which, across all model pairs, the model favours one model over the other.

$$p_{pop} = sum(1 \rightarrow N)\frac{|m_1 - m_2|}{2n_{pop}} \tag{2}$$

where N is the set of all model pairings, and n_{pop} is the total number of trials across all model pairings.

Our approach differs from several existing studies, which measure persuasion success by the difference between an agent's / model's agreement with a proposition before and after the persuasive interaction [15, 31, 32]. Given the importance of parties' arguments in legal settings, we do not consider that the pre-persuasion baseline is suitable in this context, while the availability of multiple Advocate models enables our alternative metric. (Comparing persuadability and translating metrics across bilateral and trilateral contexts will be an important avenue for future research.)

5 Results

Overall Results. Table 1 reports the Population Persuadability (p_{pop}) and Maximum Pairwise Persuadability (p_2max) of each model, in both *with arguments* and *without arguments* setups.

We first observe that, across all models and setups, p_{pop} and p_2max are statistically significant. All of our models are, to some extent, persuadable. p_{pop} ranges from 0.08 up to 0.2008, while p_2max runs from 0.1311 to 0.4052. This means that, across our full range of models, the identity of the advocate model (and hence the quality of the argument presented) has an average effect of between 8% and 21%, implying stronger Advocate models typically win between 58% and 71% of the time. As between the strongest and weakest Advocate models, depending on the Judge model, those win rates range from 63% to over 90%. We therefore conclude that all our Judge models are to some quite substantially persuadable.

As is clear from Table 1, the extent of persuadability varies substantially across Judge models. Pairwise persuadability (p_2) also varies depending on the Advocate models involved[2]. Our Advocate models were selected to ensure varying persuasiveness, so it is unsurprising to find larger p_2 where a stronger Advocate is paired with a weaker one than for two Advocates of similar strength. p_2 is typically lowest and non-significant for `claude` vs `gemini` (two similarly persuasive models) and highest (and in all bar one case significant) for `gpt4o` vs `gpt5.1` (where there is the greatest performance gap).

We observe some support for the hypothesis that larger models are less persuadable than smaller models. However the evidence here is mixed. Amongst our closed models, where we test pairs of larger and smaller models from the same model families, we see, in most cases, the larger / full model is less persuadable than the smaller / lite model. The primary exception is `gpt5-nano_minimal-reasoning`, which has the lowest p_{pop} of any model in our trial, in both the *with arguments* and *without arguments* conditions. The other exception is the marginally higher p_{pop} for `claude-sonnet` with a low thinking budget, compared to `claude-haiku` with thinking disabled. Amongst our open models, the `Qwen-8b` models appear less persuadable than `Qwen-32b`, across both thinking and non-thinking settings. `Mistral` and `Magistral` are two of the smaller models in our trial, but also amongst the least persuadable.

Similarly, we observe some support for the hypothesis that models with reasoning architecture are less persuadable than those without, particularly for larger model sizes. Amongst our large closed models, a higher reasoning setting / thinking budget corresponds to lower persuadability in five of six cases (The exception is `gemini-3-pro` in the setup without original arguments). For our smaller closed models and a number of our open models (`Qwen-32b`, `Mistral`/`Magistral`) this relationship is reversed, with higher reasoning variants appearing more persuadable on our metric. However in only one case (`gpt-5-nano` with original arguments) is this difference statistically significant at the population level[3].

We hypothesise that these unexpected results reflect the different form that persuadability takes in our experiments. Bilateral persuasion scenarios test a model's pre-interaction response, and then present it with arguments tending in one direction only, measuring how far its response changes given this persuasive treatment. In that setup a model need not be capable of evaluating arguments in order to be persuaded by them. By contrast, in our setup models are presented with competing arguments. To be persuaded by the stronger argument, a model must actually evaluate those arguments to identify which is in fact stronger. For a model that is not capable of evaluating arguments, we would expect the two competing arguments to be equally convincing, and hence lower overall persuadability. We are exploring this point in ongoing research.

Legal Substance vs Argumentative Form. We use two strategies to identify whether / how far persuadability is a function of Advocate models presenting novel arguments, that the judge model might not have otherwise considered, and hence improving the quality of the decision, versus reflecting the greater fluency or rhetorical powers of the relevant models.

First, we use two different setups to prompt our Advocates, in one of which we provide only the facts, and in the other we present also a summary of the arguments actually made in the case. We observe (see Table 1) that in 14 of 20 model setups, population persuadability is lower in the *with arguments* setup, suggesting that content, as opposed to form, is playing some role in persuasion. However, the difference is generally small, and in no case statistically significant (Chi Squared test, p<0.05).

To further test whether providing summaries of the arguments in the original case made a weaker Advocate more persuasive, we ran head-to-head trials involving the same Advocate model against itself, with one version prompted with arguments, and the other without. We thus directly test whether, given the same Advocate model, providing the original arguments increases persuasiveness. We trial this setup with two Advocates, `gpt4o` and `gpt5.1`, and four Judge models (`claude-sonnet-4.5_8k`, `deepseek-chat`, `gemini-2.5-flash-lite_reasoning`, and `gpt5.1_medium-reasoning`), with each head-to-head trial comprising 200 tests per Judge (800 tests in total).

Across all eight head-to-head trials (pairing two Advocate models across four Judges), the Advocate prompted with arguments won more frequently in every trial. However the treatment effects are small, with only one case (`gpt4o` as Advocate, `deepseek_chat` as Judge) reaching the threshold of significance ($p < 0.05$). If we aggregate across all eight trials, applying a binomial test to this 8-0 result gives a p-value of 0.0039. i.e. it seems clear that, while the effect size is small, providing original arguments when prompting our Advocates does affect persuasiveness, in turn implying that the substantive legal content of the argument, as opposed to mere form, plays at least some role.

Second, we disaggregate results by jurisdiction (US / England and Wales / Ireland) on the assumption that models have relatively greater knowledge of US law, and relatively lesser knowledge of England and Wales, and especially Irish law[4]. There is thus greater scope for persuasion based on legal content in larger jurisdictions. If we observe higher persuadability in jurisdictions where Judge models are assumed to have greater knowledge, this suggests that persuasion is driven at least in part by the quality of substantive legal arguments, as opposed to rhetorical skill or form.

For 13 of our 40 Judge model set-ups, we find a statistically significant variation across jurisdictions. Variation is in most cases in the direction which we would expect if substantive legal content explained at least part of the observed persuadability. i.e. p_{pop} is lower for our Irish scenarios than for our UK scenarios, which are in

[2]Full results including confidence intervals and p-values for all Advocate pairs are available in our data repository.

[3]Chi Squared test, p<0.05. Full significance tests are included in the data repository.

[4]See [8, 9] for discussion on varying LLM knowledge by jurisdiction.

turn lower than our US scenarios. In 28 (of 40) cases, UKp_{pop} is lower than USp_{pop}, and in 32 cases IRLp_{pop} is lower than UKp_{pop}. In 34 cases, IRLp_{pop} is lower than USp_{pop}. We thus again find suggestive evidence that the quality of substantive legal arguments, as opposed to mere rhetorical skill, plays a role in model persuadability.

6 Conclusions and Future Work

We present results showing the persuadability of a range of frontier closed and open-weights LLMs faced with hard legal questions from real world cases. We confirm these models' persuadability, show how this varies across models, and offer some explanations for why this may be the case. We also offer some tentative conclusions on how far persuadability is a function of the content versus the form of arguments presented.

Our motivating question was whether and to what extent LLMs meet the requirement that an administrative or judicial decision-maker is capable of being persuaded, while also being able to make and stand over their own decisions. In this light, some conclusions can be drawn. In the case of the smaller models, where lower persuadability may reflect difficulty evaluating competing arguments, our results suggest these models are to that extent inappropriate in this role. In contrast, the lower persuadability of our larger models is likely better explained by these models forming their own views about the substantive question. However, even for these larger models, very high figures for p_2max in particular indicate that they are, at least sometimes, very strongly affected in their decisions by the form and content of arguments presented to them. Whether this is excessive is ultimately a political question, reflecting our expectations of our justice system, and how it treats less able subjects in particular. At a minimum, these choices should be made cognisant of the persuadability characteristics of candidate models.

Our research points towards a number of open questions. First, which features are persuading our Judge models in particular instances? Second, does (and under what circumstances) exposure to argument improve the quality of decisions reached? And third, how does persuadability of models in relation to hard questions compare with relevant human experts, including senior lawyers and judges? While the resource implications of this latter task are substantial, we hope that future research can take up these tasks.

References

[1] Guilherme F.C.F. Almeida et al. 2024. Exploring the Psychology of LLMs' Moral and Legal Reasoning. *Artificial Intelligence* 333 (Aug. 2024), 104145. doi:10.1016/j.artint.2024.104145

[2] Amalia Amaya. 2025. Reasoning in Character: Virtue, Legal Argumentation, and Judicial Ethics. *Ethic Theory Moral Prac* 28, 3 (July 2025), 359–378. doi:10.1007/s10677-023-10414-z

[3] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can GPT-3 Perform Statutory Reasoning?. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23)*. Association for Computing Machinery, New York, NY, USA, 22–31. doi:10.1145/3594536.3595163

[4] Andrew Blair-Stanek and Benjamin Van Durme. 2026. LLMs Provide Unstable Answers to Legal Questions. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (ICAIL '25)*. Association for Computing Machinery, New York, NY, USA, 425–429. doi:10.1145/3769126.3769245

[5] Simon Martin Breum et al. 2024. The Persuasive Power of Large Language Models. *ICWSM* 18 (May 2024), 152–163. doi:10.1609/icwsm.v18i1.31304

[6] Carlos Carrasco-Farre. 2024. Large Language Models Are as Persuasive as Humans, but How? About the Cognitive Effort and Moral-Emotional Language of LLM Arguments. arXiv:2404.09329 [cs]

[7] Andrew Coan and Harry Surden. 2025. Artificial Intelligence and Constitutional Interpretation. *U. Colo. L. Rev.* 96, 2 (2025), 413–498.

[8] Damian Curran et al. 2025. Place Matters: Comparing LLM Hallucination Rates for Place-Based Legal Queries. arXiv:2511.06700 [cs]

[9] Matthew Dahl et al. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (Jan. 2024), 64–93. doi:10.1093/jla/laae003

[10] Esin Durmus et al. 2024-04-09, 2024. Measuring the Persuasiveness of Language Models. https://www.anthropic.com/news/measuring-model-persuasiveness

[11] John Gardner. 2001. Legal Positivism: 5 1/2 Myths. *Am. J. Juris.* 46 (2001), 199.

[12] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models. *Proc. Natl. Acad. Sci. U.S.A.* 121, 24 (June 2024), e2403116121. doi:10.1073/pnas.2403116121

[13] Mateusz Idziejczak et al. 2025. Among Them: A Game-Based Framework for Assessing Persuasion Capabilities of LLMs. In *Advances in Knowledge Discovery and Data Mining*, Xintao Wu, Myra Spiliopoulou, Can Wang, Vipin Kumar, Longbing Cao, Yanqiu Wu, Yu Yao, and Zhangkai Wu (Eds.). Vol. 15874. Springer Nature Singapore, Singapore, 183–195. doi:10.1007/978-981-96-8186-0_15

[14] Tianjie Ju et al. 2025. On the Adaptive Psychological Persuasion of Large Language Models. doi:10.48550/arXiv.2506.06800

[15] Shirish Karande, Santhosh V, and Yash Bhatia. 2024. Persuasion Games with Large Language Models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. NLP Association of India (NLPAI), Chennai, India, 576–582. https://aclanthology.org/2024.icon-1.67/

[16] John M Kelly. 1964. Audi Alteram Partem;Note. *NATURAL LAW FORUM* (1964).

[17] Jinqi Lai et al. 2024. Large Language Models in Law: A Survey. *AI Open* 5 (2024), 181–196. doi:10.1016/j.aiopen.2024.09.002

[18] José Luiz Nunes, Guilherme Almeida, and Brian Flanagan. 2025. Evidence of Conceptual Mastery in the Application of Rules by Large Language Models. doi:10.2139/ssrn.5161877

[19] OpenAI. 2024. OpenAI O1 System Card. doi:10.48550/arXiv.2412.16720

[20] Paulina Jo Pesch. 2025. Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making. *Eur. j. risk regul.* 16, 1 (March 2025), 76–95. doi:10.1017/err.2024.99

[21] Eric A. Posner and Shivam Saran. 2025. Judge AI: Assessing Large Language Models in Judicial Decision-Making. social science research network:5098708 doi:10.2139/ssrn.5098708

[22] Alexander Rogiers et al. 2024. Persuasion with Large Language Models: A Survey. arXiv:2411.06837 [cs]

[23] Philipp Schoenegger et al. 2025. Large Language Models Are More Persuasive Than Incentivized Human Persuaders. arXiv:2505.09662 [cs]

[24] Lawrence B. Solum. 2003. Virtue Jurisprudence A Virtue–Centred Theory of Judging. *Metaphilosophy* 34, 1-2 (Jan. 2003), 178–213. doi:10.1111/1467-9973.00268

[25] Bryan Chen Zhengyu Tan et al. 2025. Persuasion Dynamics in LLMs: Investigating Robustness and Adaptability in Knowledge and Safety with DuET-PD. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 1550–1575. doi:10.18653/v1/2025.emnlp-main.81

[26] Cassandra Teigen et al. 2024. Persuasiveness of Arguments with AI-source Labels. *Proceedings of the Annual Meeting of the Cognitive Science Society* 46, 0 (2024). https://escholarship.org/uc/item/6t82g70v

[27] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Moral Alignment for LLM Agents. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=MeGDmZjUXy

[28] Jeremy Waldron. 2023. The Rule of Law. In *The Stanford Encyclopedia of Philosophy* (fall 2023 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/rule-of-law/

[29] Yi Zeng et al. 2024. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 14322–14350. doi:10.18653/v1/2024.acl-long.773

[30] Kepu Zhang et al. 2025. SyLeR: A Framework for Explicit Syllogistic Legal Reasoning in Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, Seoul Republic of Korea, 4117–4127. doi:10.1145/3746252.3761120

[31] Haodong Zhao et al. 2025. Disagreements in Reasoning: How a Model's Thinking Process Dictates Persuasion in Multi-Agent Systems. arXiv:2509.21054 [cs]

[32] Xiaochen Zhu et al. 2025. Conformity in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 3854–3872. doi:10.18653/v1/2025.acl-long.195