Too Fast, Too Shallow – LLMs, Including Reasoning LLMs, Are Unreliable Constitutional Reasoners

Reto Gubelmann
University of Zurich
reto.gubelmann@uzh.ch

Peter Hongler
University of St. Gallen
peter.hongler@unisg.ch

Abstract

We assess LLMs' constitutional reasoning abilities using three different, newly developed datasets on three different constitutional questions in three different constitutional frameworks, comprising two different languages; the structure and content of the datasets is informed by legal expertise and grounded in the state of the art in philosophy of language. Our results indicate that the 19 LLMs tested, including the reasoning LLMs, while not being uniformly subject to political bias, are still not reliable constitutional reasoners, as they are heavily influenced by logically irrelevant aspects of the reasoning. Of the 196k evaluations run in our main experiment, the LLMs label less than 70% correctly, and open-weight reasoning LLMs as well as gpt-4o are outperformed by moderately sized open-weight non-reasoning LLMs. None of the LLMs tested consistently displayed slow, deep, rule-based system 2 thinking.

1 Introduction

The dream of building an AI system that is capable of reasoning in a way that is less subjective or biased than a human being is an old one. In the legal domain, there exists the idea of formalism, an algorithmic, mathematical procedure of judging that leaves no room for bias or subjectivity to deduce an unbiased, objective judgment from a wealth of data and perfect knowledge of the law.[1] With the advent of LLMs and their reasoning abilities, some scholars think that the formalist's vision might now finally become reality (Buckland, 2023). Our study aims to contribute to clarifying whether LLMs are indeed able to realize the formalist's dream.

We conceptualize this research goal using the influential dual process theory of cognition proposed by Tversky and Kahneman (1974); Kahneman (2003); Kahneman et al. (2021). According to the theory, humans have two different processes, or systems of reasoning about a subject matter. The first one is the older one, it is fast and associative, relying on heuristics and biases. The second one, in contrast, is evolutionarily more recent, much slower, and it is governed by explicit rules, including logical rules of inference (see also Kogler and Kühberger 2007; Evans and Curtis-Holmes 2005; Frankish 2010). While system 1 thinking has served us well and continues to serve us well in situations where quick decision and action are essential, we would like our artificial judges to rely on system 2 rather than system 1 reasoning. Judges that rule based on deep understanding of the relevant conceptual, argumentative and logical relations are preferable to judges ruling based on a shallow understanding of these relations, on heuristic rules of thumb, and on various kinds of bias (for more details on the concepts of shallow heuristics and bias as used here, see Gubelmann et al. 2022). If it would turn out that LLMs judge in a system 1 manner, then they might not be able to realize the formalist ideal any more than humans.

To operationalize our research goal conceptualized in this way, we measure the impact of two

Please check the following reasoning for its deductive-formal validity. Ignore any content of the inference and only focus on its form, ending your answer with either **valid** or **invalid**: The US constitution implies no right to abortion [C], because the right of privacy, as implied by the concept of liberty in the fourteenth amendment of the US constitution, is compatible with a complete ban on abortion [P1], because the remainder of the US constitution and all of its amendments are compatible with a complete ban on abortion [P2], because the sun rises every day [P3], and because basic considerations of human dignity are compatible with a complete ban on abortion [P4].

Table 1: A sample from our dataset. The Conclusion C is deductively entailed by premises P1 and P2, while P3 and P4 (the random premise "rand" and an important-sounding but irrelevant reference to human dignity "HD") do not help, but also do not hurt this validity. In our experiment, none of our 19 LLMs correctly labelled this inference in any of the three runs we conducted.

[1] This idea can be traced back to Kelsen (2017–1934).

Findings of the Association for Computational Linguistics: ACL 2026, pages 40554–40572

factors on how LLMs label the validity or invalidity of constitutional arguments: (1) the political leaning of the conclusion of the argument to be judged, testing the political bias of LLMs, and (2) the sensitivity of LLM performance to the presence of specific logically irrelevant perturbations, assessing whether LLMs rely on shallow heuristics rather than on an actual understanding of logical concepts.

We synthetically build a triplet of datasets consisting of arguments in three different constitutional traditions: the discussion surrounding a constitutional right to abortion in the US, the discussion surrounding a human rights obligation of states to reduce greenhouse gas emissions in Europe,[2] and the discussion surrounding the constitutional permissibility of regressive income taxes in Switzerland. To measure bias, we construct the dataset such that each argument favoring one side of the debate has a twin that is identical except for supporting the other side of the relevant debate. To measure reliance on heuristics, we systematically vary different factors to control for the influence of such aspects as negation, premise and conclusion order, or the presence of a random premise.

On this triplet, we assess a total of 19 different LLMs, both open-weight and closed-source, in a number of different instruction settings, including three different few-shot settings. An example of a fully specified prompt can be seen on Table 1; notably, it is a sample that none of the LLMs labelled correctly. There, the political leaning of the inference is given by the content of both the premises printed in green (P1, P2) and conclusion (C) printed in olive, which bend towards the conservative position on the issue, and the perturbation is given by a patently irrelevant, but logically harmless additional premise in red (P3) as well as relevant-sounding, but equally logically irrelevant and harmless premise invoking human dignity printed in maroon (P4). We assess LLMs' ability to accurately label the validity or invalidity of these arguments.

Our study makes two contributions. First, we develop a legally sound, topically, culturally, and linguistically diverse set of three synthetic, high-quality datasets focusing on legal reasoning on US, European, and Swiss constitutional matters (comprising two languages: English and German) grounded in recent developments in the philosophy of language and logic, which are publicly available.[3] Second, we rely on this dataset to conduct comprehensive experiments involving 19 open- and closed-source LLMs using different instruction paradigms, finding that current reasoning (Ke et al., 2025) and non-reasoning LLMs exhibit strong signs of reliance on shallow heuristics, that is, of system 1 reasoning.

Addressing this topic is important because constitutional reasoning is of central societal importance, potentially impacting the rights and duties of millions of people; beyond legal reasoning, any use of LLMs in decision-making is put to question if evidence of fast, shallow system 1 thinking accumulates.

2 State of Relevant Research

Material Inference The concept of a material inference has been pioneered in recent philosophy of language by Robert Brandom (Brandom, 1994, 2010, 2021). Before Brandom, philosophy of language focused on formal logic, that is, the kind of inference on display in example (1) (see, e.g., the influential work by Quine 1960, 1974, 1995).

(1) All house cats have a black belt in Karate, and all US Supreme Court Judges are house cats. Therefore, all US Supreme Court Judges have a black belt in Karate.

There cannot be a rational dispute regarding the validity of (1): If the premises are true, then the conclusion must necessarily be true as well. We emphasize that *validity* must be distinguished from *soundness*: While a valid inference can have premises that are patently false (like example (1)), an inference is sound only if it is valid *and* if its premises are true. For an example of a material inference, consider example (2).

(2) The constitutional right to privacy entails a right to abortion.

In contrast to Example (1), there can indeed be a reasonable disagreement regarding the validity of inference (2). Whether it is valid depends on one's understanding of the constitutional right to privacy – as the recent ruling by the Supreme Court of the US has shown, and as decades of arguments by legal scholars have shown, there can be rational disagreement about that. In our experiments, we combine material and formal inferences in a way

[2] Technically, this second case relates to the interpretation of the European Convention on Human Rights; however, being a human rights convention, it has constitutional features.

[3] https://github.com/retoj/llms-constitutional-reasoning.

that should reveal reliance on bias rather than an understanding of logical concepts: As on display in Table 1, **the premises of our samples contain politically contested material inferences; the entire argument, however, constitutes a formally valid inference whose validity, like Example (1), does not depend on political predilections.**

NLI Contemporary Natural Language Inference (NLI) is conceived as a three-way classification problem: given two claims, their logical relation must be classified as either entailment, contradiction, or neutral. (see Dagan et al. 2006; Bar Haim et al. 2006; Giampiccolo et al. 2007; Bentivogli et al. 2009, 2011; Gubelmann et al. 2024). Encoder-Only Transformers such as BERT (Devlin et al., 2019) have been shown to succumb to the so-called Problem of Generalization (Bernardy and Chatzikyriakidis, 2019; He et al., 2019; Mahabadi et al., 2019; Zhou and Bansal, 2020; Bras et al., 2020; Utama et al., 2020; Asael et al., 2021; Gubelmann et al., 2022, 2023). Rather than developing an actual understanding of logical relationships that can generalize beyond the respective datasets, the models were relying on shallow heuristics that perform well on the benchmarks because of artifacts in these benchmarks, but which are useless in the wild. For example, Gururangan et al. (2018) found a so-called negation bias on which these models were relying (on negation, see also Gubelmann and Handschuh 2022). In sum, this indicates that the encoder-only transformers were relying on system 1 reasoning instead of system 2 reasoning.

However, it has been hypothesized that the shallow heuristics were induced by the fine-tuning step on the NLI-datasets, which contained artifacts. In the current paradigm of decoder-only LLMs, the fine-tuning step is often less important. Therefore, it is an open question whether contemporary generative LLMs would still show such behavior. The research regarding their inferential abilities is still evolving. Liu et al. (2023) find generally good performance, Payandeh et al. (2023) discover some brittleness. We contribute to this ongoing research effort by testing a large number of LLM configurations in the challenging arena of constitutional reasoning and inference.

Political Bias in LLMs & Dual Process Theory Research on political bias in LLMs is still evolving. The available studies suggest that the LLMs have a left-wing-bias, see Ceron et al. (2024) and Motoki et al. (2024). Regarding system 1 and system 2 thinking in LLMs, there are a number of studies that examine specific biases, such as anchoring, halo, overconfidence, or recency bias, finding mixed results overall: In some settings, the LLMs show strong bias indicative of system 1 thinking, in others, evidence of system 2 reasoning was found (Echterhoff et al., 2024; Hagendorff et al., 2023; Ziabari et al., 2025). However, system 1 vs. system 2 thinking has never been explored in direct connection with logical inferences. We fill this gap.

Within dual process theory, our assessment of political bias via inference labelling behavior belongs to the research strand focusing on the logic-belief conflict to distinguish between system 1 and system 2 reasoning: While system 2 thinking shows itself in its ability to think hypothetically and rule-based, system 1 thinking is quick and associative and tends to confuse perceived truth of premises and conclusions with validity of inference (Evans et al., 1983; Evans and Curtis-Holmes, 2005; Roberts and Newton, 2011). Our contribution combines interest in political bias with the notion of material inference and applies both to the societally highly relevant case of constitutional reasoning.

3 Datasets

3.1 General Structure of Datasets

Political Bias We start out with basic, formal-deductively valid inference patterns that defend a position that we label "left" and "right" respectively. We hold that our labelling roughly aligns with current understanding of the political landscape (see below, Section 3.2); **we emphasize, however, that nothing in our argument depends on these labels**. We also emphasize that, while there is intentional and systematical partiality regarding the material inferences within the premises and conclusions, the inferences as a whole are all formally valid or invalid regardless of any political position.

Shallow Heuristics We then introduce five variations of each of these basic, right- and left-leaning arguments to measure reliance on various kinds of shallow heuristics. **Validity**. We wanted to measure whether the LLMs rely on some sort of validity heuristics, e.g., simply labelling valid all the time. Therefore, we paired each of the valid arguments with an obviously invalid one (basically by combining right-leaning premises with left-leaning conclusions and *vice versa*, see Table 2 for an illustration). **Negation**. "prem-neg?" and "con-neg?"

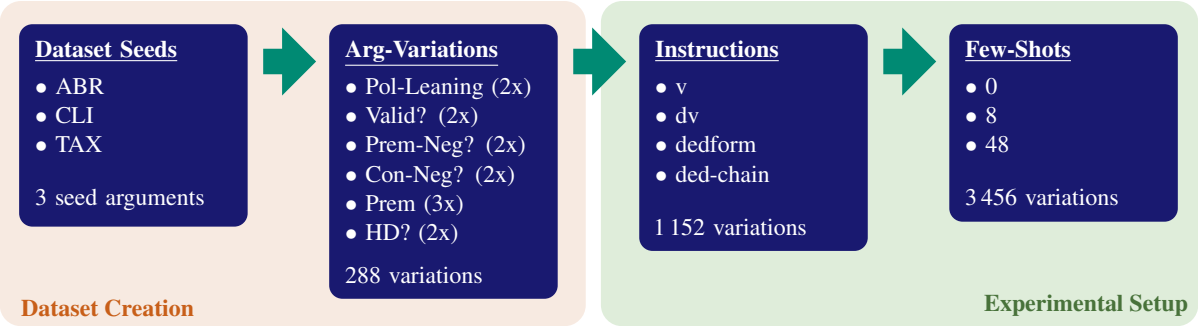

Figure 1: Overall sample creation pipeline. Multiplied by 19 LLMs times 3 runs, this results in the 196 992 inferences run in our experiment.

control whether the conclusion or the premise contain a negation. Earlier research (see above, Section (2)) has shown that Encoder-Only transformers rely on various kinds of negation heuristics, so we wanted to ensure that our patterns are not only balanced regarding negation, but that we are also able to test and examine the influence of different combinations of negated and positive premises. **Modification of Premises** "Prem" refers to three variations of the basic premise pattern, namely the default one, another one (called "conlast") where the conclusion comes at the end, and one with the introduction of a random sentence (called "rand", "the sun rises every day", on display in red in Table 1). **Important-Sounding But Irrelevant Addition** "HD?" controls whether an additional, highly schematic sentence involving human dignity (for ABR and CLI) or Tax Justice (for TAX) is added. We wanted to see whether the models are distracted by this largely vacuous, superficially relevant, but also in fact useless and harmless insertion of an additional premise (illustrated in maroon on Table 1).

Figure 1 gives an overview on how we construct our datasets to test for political bias and reliance on various kinds of shallow heuristics. According to it, we receive per initial argument variations in pol-leaning, con-neg?, prem-neg?, Valid?, Prem, and HD?, yielding 96 variations in total per dataset. The three datasets (see the next section) then yield a total number of 288 argument samples.

3.2 The Three Datasets

Based on the structure on display in Fig 1, we compile three different datasets from three different topical fields, each connected to a topical matter that's politically contested. All of the three datasets are based on real court cases and had been suggested by the legal expert in our team. However, every argument of our dataset is synthetically constructed by humans or automatically generated by simple syntactical patterns out of human-generated ones. We emphasize that our dataset spans three different constitutional traditions, equally many legal domains, and two different languages (English and German). While we believe that this way of constructing our dataset allows us to strike a good balance between diversity and control, we think that the highly schematic nature of our datasets might yield results that are higher than what the LLMs could achieve in the wild.

"ABR": US; Does the right to Privacy Entail a Right to Abortion? In the Supreme Court decision "Dobbs v. Jackson Women's Health Organization", 597 U.S. 215 (2022), the US Supreme Court ruled that the US Constitution does not imply a right to abortion. The core question at issue was whether the right to privacy, as implied by the fourteenth amendment, in turn implies a right to abortion. Proponents of a constitutional right to abortion typically claim that it implies a right to abortion, while opponents, including the 6-3-majority of the current US Supreme Court, think otherwise.

"CLI": EU; Does the Right to Private Life Entail an Obligation of States to Reduce GHG Emissions? The central question here is whether the right to private and family life, according to Art. 8 of the Convention for the Protection of Human Rights and Fundamental Freedoms, implies an obligation on states to reduce their greenhouse gas emissions. In a ruling from April 2024, the European Court of Human Rights ruled that it does, following the argument of the plaintiffs "Verein Klimaseniorinnen" (literal translation: "Association of Climate Elderly Women") from Switzerland. The ruling has immediately caused a very lively debate. Experts, including a former Supreme Court judge, have criticized the ruling harshly, claiming that the

Conclusion	Argument Patterns
The US constitution is compatible with a complete ban on abortion *The US constitution implies a right to abortion*	the right of privacy, as implied by the concept of liberty in the fourteenth amendment of the US constitution, is compatible with a complete ban on *implies a right to* abortion; the remainder of the US constitution and all of its amendments are compatible with a complete ban on abortion
The right to respect private and family life in art 8 of the ECHR *requires that each contracting state undertake measures for the substantial and progressive reduction of their respective* is consistent with contracting states' remaining inactive regarding their respective GHG emission level	the right to respect private and family life in Art. 8 of the European Convention on Human Rights (ECHR) contains the obligations of states to protect individuals from adverse effects on human health; this protection from adverse effects on human health includes the protection from various sources of environmental harm; this obligation to protect individuals from various sources of environmental harm *requires* does not require that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level; no other obligation can be derived from article 8 of the ECHR to require that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level
degressive Einkommenssteuern sind *verfassungswidrig* verfassungskonform	*die Schweizer Bundesverfassung enthält ein Leistungsfähigkeitsprinzip, gemäss der sich die Besteuerung nach der wirtschaftlichen Leistungsfähigkeit richten muss; das Leistungsfähigkeitsprinzip schliesst degressive Einkommenssteuern aus*; das Leistungsfähigkeitsprinzip der Schweizer Bundesverfassung lässt degressive Einkommenssteuern zu; alle übrigen Bestimmungen der Schweizer Bundesverfassung lassen degressive Einkommenssteuern zu

Table 2: Sample argument patterns of the ABR, CLI, and TAX datasets. Combining parts underlined and in red (italicized and in blue) yields a valid right-leaning (left-leaning) argument, combining blue premises with red conclusions and *vice versa* yields invalid inferences.

judges have created new law rather than merely interpreting existing law.[4] Our CLI dataset is built around this question of whether an obligation to reduce GHG emissions can be derived from the protection of private and family life.

"TAX": CH; Does the Taxation Principle of the Ability to Pay Prohibit Regressive Income Taxes? It was disputed whether the ability to pay principle in Art. 127 para. 2 of the Swiss Constitution limits the rights of Swiss cantons to implement regressive (in German: "degressive") income tax rates. It is, therefore, essentially contested whether the ability to pay principle as a constitutional principle contains such a limitation. Regressive tax rates mean that the average income tax rate decreases after a certain amount of income. This was the case in the canton of Obwalden at the time of the decision. The court concluded (see BGE 133 I 206) that the ability to pay principle is neither compatible with regressive income taxes, nor with regressive wealth taxes. The canton had to change its tax laws, which had previously been adopted by a majority of the canton's citizens through a popular vote. We display samples of formally valid and invalid inferences for all of these three datasets in Table 2 (for further samples, see the Appendix, Tables 7 and 8).

Quality Assurance All arguments were co-developed with a legal scholar specializing in constitutional questions and checked by a philosopher of language specializing in inferentialism. In case of disagreement, the sample was modified to fit both expectations. This means that the dataset was developed in an inter-disciplinary, co-creative process, rather than by annotating pre-existing data. Therefore, it is not helpful to give data on inter-annotator agreement, such as Cohen's Kappa (Warrens, 2015): As a consequence of the co-creative generation process, all annotators entirely agree on the categorization of arguments.

4 Experiment

Models For our experiments, we used a total of 19 LLM configurations, including small and large versions from llama3/llama3.3 and gemma2, mistral, mixtral-8x7b, all in two different precisions, two sizes (and precisions) of distillations of the pioneering reasoning LLM deepseek-r1, as well as gpt-4o from OpenAI. We add two more open LLMs with full precision that have been optimized for German text (recognizable by the qualifier "sauerk" in the results). We run the entire experiment three times and average over the results of the three runs. For the technical details of our experiment, see the appendix, Section B.

Instructions We start with four different instructions, see Table 3, which we combine with three different few-shot settings (0, 8, and 48 few-shots), yielding a total of 12 different instruction settings.

[4] See this report, accessed on September 2, 2024.

Label	Instruction
v	"Please check whether the following reasoning is valid, ending your answer with either **valid** or **invalid**"
dv	"Please check whether the following reasoning is deductively valid, ending your answer with either **valid** or **invalid**"
dedform	"Please check the following reasoning for its deductive-formal validity. Ignore any content of the inference and only focus on its form, ending your answer with either **valid** or **invalid**"
ded-chain	"Please check whether the following inference is deductively valid; this means that you have to assess whether it is possible that the premises are true, while the conclusions are still false. If this is possible, it is invalid, otherwise, it is valid. End your answer with either **valid** or **invalid**"

Table 3: Our four versions of instructions with increasing thrust towards system 2 thinking.

We balance the few-shots for political leaning and validity, and we take them from a different (US) constitutional topic, namely the right to bear arms, in order to avoid biasing the LLMs towards one of our three datasets. For details of these patterns, see the Appendix, Section C.

While *v* is rather implicit, the phrasing becomes ever more explicit until, with *ded-chain*, we give the entire definition of validity together with a procedure on how to assess it. With humans, this should provoke system 2 thinking (see Evans 2003).

This yields a total number of 1 152 prompts per dataset (96 variations of arguments multiplied by 12 different instructions). Hence, in total, we test the models on 3 456 samples from three datasets, which we run three times, meaning that we have 10 368 inferences in total per LLM; given the 19 LLMs tested, we ran a total of 196 992 inferences.

Postprocessing As we did not want to restrict the natural flow of generation of the LLMs, we only requested that the validity label be produced following a certain pattern (see Table 3), but we did not restrict the length or shape of the generated output in any other form. This means that we had to extract the labels or grades from this output. We did so in a postprocessing step. Postprocessing occurred with a set of hand-crafted regex-patterns to ensure that the correct label is extracted from the (mostly non-restricted) output of the LLMs. Postprocessing worked generally well, only with three LLMs falling below 95% recall, and never below 90%. Details on the samples discarded during postprocessing (because no label could be extracted) can be found in the Appendix, Table 13; manual inspection of 200 extracted labels showed that precision of extraction is above 99%.

	8 FS			48 FS		
Modelname	inv	val	Mean	inv	val	Mean
llama3.3-70B:4b	0.48	0.95	0.71	0.97	0.71	**0.84**
llama3.3-70B:16b	0.58	0.93	0.76	0.87	0.76	**0.82**
mixtral-8x7B:16b	0.95	0.63	**0.79**	0.90	0.57	0.74
gemma-sauerk	0.98	0.58	**0.78**	1.00	0.38	0.69
gemma2-9B:32b	0.99	0.57	**0.78**	1.00	0.38	0.69
gemma2-9B:4b	0.99	0.57	**0.78**	1.00	0.34	0.67
mixtral-8x7B:4b	0.93	0.63	**0.78**	0.90	0.57	0.73
gpt-4o	0.77	0.76	**0.77**	0.84	0.61	0.72
gemma2-27B:4b	0.98	0.51	**0.75**	0.99	0.35	0.67
ds-r1-70B:4b	0.80	0.68	**0.74**	0.83	0.64	0.73
gemma2-27B:32b	0.99	0.49	**0.74**	0.99	0.34	0.67
mistral-7B:4b	0.56	0.78	0.67	0.95	0.36	0.66
mistral-7B:16b	0.59	0.77	**0.68**	0.97	0.32	0.65
llama-sauerk	0.91	0.44	**0.68**	0.99	0.12	0.55
ds-r1-70B:16b	0.58	0.68	0.63	0.57	0.75	**0.66**
llama-8B:16b	0.77	0.53	**0.65**	0.88	0.32	0.60
llama3-8B:4b	0.73	0.52	**0.63**	0.88	0.27	0.58
ds-r1-7B:16b	0.69	0.53	**0.61**	0.75	0.39	0.57
ds-r1-7B:4b	0.66	0.50	**0.58**	0.69	0.40	0.54
Mean Of Models	0.79	0.63	0.71	0.89	0.45	0.67

Table 4: Mean accuracy by model, few-shots-setting (FS) and validity, with bolded maximum values and sorted by highest mean. Not shown are results for 0 FS, as they are, with one single exception (mistral 4b, 0.68 instead of 0.67) inferior to the other FS settings.

5 Results

Table 4 shows that, overall, accuracy varies substantially with the lowest at 0.58 and thus little above chance, while the best accuracy is at 0.84, achieved by llama3.3 with 48 few-shots. This table also shows that, overall, the best few-shot setting is 8 FS. However, both the quantized and the full-precision version of llama3.3 benefit substantially from 48 few-shots. This is also true, albeit on a much lower level, for the full-precision version of deepseek-r1-70B. While the number of few-shots has a clear impact on model behavior and performance, the specific instruction used in combination with the few-shots has moderate effects on model behavior. We show the corresponding numbers in the Appendix, Table 12.

Figure 2 compares the average accuracies of reasoning vs. non-reasoning LLMs, subdivided by dataset. To give a comprehensive overview, this figure also already includes the results from our ICE ablation dataset, which we introduce below, Section 6. This figure shows a substantially different behavior of the two kinds of LLMs: e.g., non-

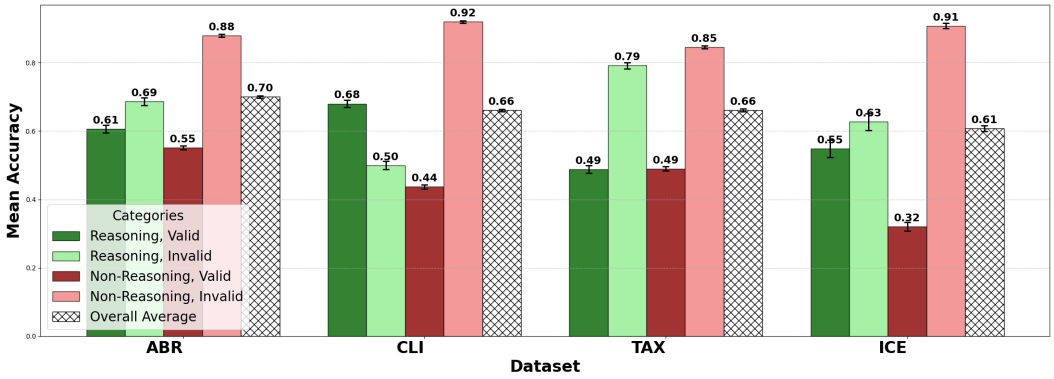

Figure 2: Mean accuracy of reasoning vs. non-reasoning LLMs, by dataset and validity with 95% confidence.

reasoning LLMs perform much better with invalid inferences across all three datasets, while reasoning LLMs only do so clearly with the German TAX dataset and slightly with the ABR dataset. Also, non-reasoning LLMs perform better than reasoning LLMs overall: average accuracy across all FS settings is at **0.69** for non-reasoning LLMs and **0.63** for reasoning-LLMs.

Table 5 shows the influence of five variables used to probe LLMs for bias or heuristics, and hence for system 1 thinking (see above, Section 3). We here filtered for the instructions with 8 FS, the strongest setting overall. **Pol.** is the relevance of political leaning, with positive numbers indicating that LLMs are more accurate at labelling left-leaning inferences than right-leaning inferences and negative numbers indicating the same for right-leaning inferences. It can be seen that the influence of political leaning is quite moderate overall, with a few outliers with significant bias, notably the popular deepseek-r1 reasoning LLMs. **Prem.** gives the difference in accuracy between the best-performing conlast and the worst-performing rand condition (see above, Section 3). **Valid** is the relevance of instruction, with positive numbers indicating that the LLM is more accurate at labelling valid inferences, while the negative numbers indicate the converse. **p-neg**, **c-neg**, and **HD** indicate whether a presence of negation in premise or conclusion or of an additional important-sounding premise increases or decreases accuracy with a model. The means given are the means of absolute differences in accuracy. Overall, it shows that the LLMs are highly sensitive to the random perturbation, and that they are considerably better at labelling either valid or, which is more common, invalid inferences. Looking at Table 4 and Figure 2, it shows that non-reasoning LLMs simply tend to label inferences as invalid,

Modelname	Pol.	Prem.	Valid	p-neg	HD	c-neg
ds-r1-70B:16b	0.10	*0.36*	0.10	0.04	0.03	-0.01
ds-r1-70B:4b	*0.14*	*0.35*	*-0.12*	0.05	-0.02	-0.01
ds-r1-7B:16b	0.06	0.03	*-0.16*	0.05	0.04	0.01
ds-r1-7B:4b	-0.04	0.04	*-0.16*	0.03	0.09	0.00
gemma2-27B:32b	-0.05	***0.82***	*-0.50*	0.08	-0.09	0.05
gemma2-27B:4b	-0.07	***0.81***	*-0.48*	0.06	-0.09	0.05
gemma2-9B:32b	0.02	***0.99***	*-0.42*	0.02	-0.04	0.03
gemma2-9B:4b	-0.01	***0.97***	*-0.43*	0.02	-0.06	0.02
gemma-sauerk	-0.01	***0.96***	-0.08	0.02	0.06	*-0.40*
gpt-4o	0.04	*0.48*	-0.00	-0.01	-0.08	0.02
llama-8B:16b	-0.02	***0.69***	*-0.23*	0.06	-0.01	0.03
llama3-8B:4b	-0.02	***0.68***	*-0.21*	0.07	-0.02	0.04
llama3.3-70B:16b	0.07	*0.18*	*0.35*	-0.00	0.02	0.01
llama3.3-70B:4b	-0.00	*0.15*	*0.47*	-0.00	0.02	0.02
llama-sauerk	-0.06	***0.74***	-0.06	0.09	*0.12*	*-0.47*
mistral-7B:16b	0.00	***0.56***	*0.18*	-0.00	0.02	0.02
mistral-7B:4b	0.03	***0.52***	*0.22*	0.05	0.01	-0.00
mixtral-8x7B:16b	0.03	***0.90***	*-0.33*	0.04	0.05	0.02
mixtral-8x7B:4b	0.03	***0.87***	*-0.31*	0.07	0.01	-0.00
Abs. Mean	0.04	***0.59***	*0.29*	0.04	0.04	0.03

Table 5: Differences in accuracy brought out by our bias and heuristic detection variations, filtered for 8 FS, the strongest setting found. Differences over 10% are italicized, over 50% are boldfaced.

especially with the CLI and TAX datasets, leading to a large difference in accuracy between the two conditions. With reasoning LLMs, this effect is less pronounced.

6 Discussion

In this discussion, we focus on the implications of the results of our experiments regarding the main research interest, as outlined above, Section 1: The question whether current LLMs are able to transcend bias and shallow heuristics, achieve genuine system 2 thinking, and hence realize the formalist's dream in constitutional reasoning. Before doing so, we make some observations regarding the three datasets. Figure 2 shows that, overall, the ABR dataset is easiest for the LLMs (mean accuracy of

Conclusion	Argument Patterns
The general inclination of U.S. citizens to eat ice cream, as evident in the sales figures of U.S. ice cream companies, is consistent with U.S. citizens' wanting to ban ice cream from clothing stores The general inclination of U.S. citizens to eat ice cream, as evident in the sales figures of U.S. ice cream companies, suggests that U.S. citizens like to eat ice cream in clothing stores	The general inclination of U.S. citizens to eat ice cream, as evident in the sales figures of U.S. ice cream companies, includes the inclinations to eat ice cream outside of their homes or appartments. This inclination *is consistent with U.S. citizens' wanting to ban ice cream from some places, such as clothing stores. All other known general inclination of U.S. citizens is consistent with U.S. citizens' wanting to ban ice cream from clothing stores* does imply the inclination to eat ice cream in all places, such as clothing stores; all other known general inclinations of U.S. citizens imply that U.S. citizens like to eat ice cream in clothing stores.

Table 6: Basic Argument Patterns of the ICE ablation dataset. Combining the parts italicized and light brown (underlined and pink) yields a valid anti-ice-cream-in-clothing-stores (pro-ice-cream-in-clothing-stores) argument, combining brown premise elements with pink conclusions yields an invalid anti-ice-cream inference, and *vice versa.*

0.7), with the CLI and TAX dataset roughly tied (mean accuracy of 0.66). We hypothesize that this is because the US Supreme Court's ruling over Roe v. Wade has been extensively discussed in the public. The climate ruling, in contrast, on which the CLI dataset focuses, is more recent. Furthermore, we emphasize that the relatively good performance with the German TAX dataset is put in perspective by the fact that both reasoning and non-reasoning LLMs show very low accuracy with valid inferences (0.49), which means that they simply tend to label samples from this German dataset invalid in general.

A factor contributing to the difficulty LLMs show with CLI could also be the complexity of the reasoning, which is typical for the legal culture of the European Court of Human Rights and particularly pronounced in the reasoning adopted by the court in this case. This also leads to longer overall arguments. While the average length of the shortest prompt configuration (instruction "v" from Table 3 and 0 FS) is 458 characters if ABR and 488 with TAX, it is 1013 with CLI (for more details on this, compare the Appendix, Section A). We note that the sheer length should not be a fundamental challenge for system 2 reasoners: humans might need a scratchpad to work through the reasoning, but it is not in principle more difficult.

6.1 Political Orientation With Little Causal Influence

Overall Mild Political Bias As Table 5 shows, overall, LLMs show mild political bias. For most LLMs, when compared to our probes for shallow heuristics, the political orientation does not have a noteworthy effect on LLM performance. We hypothesize that the rather small relevance of political leaning is a result of LLM alignment, incentivizing LLMs to refrain from taking sides in politically charged debates. Interestingly, it is the two configurations of the largest reasoning LLM tested, namely ds-r1-70B, that lose substantial accuracy, 10 and 14 percentage points respectively, when asked to label right-leaning as opposed to left-leaning arguments.

Non-Political Ablation Study: ICE However, it might be that the political nature of the inferences as such has an influence on LLM behavior. To explore whether this is the case, that is, whether the behavior of LLMs that we observe is due to the subject matter of the inferences and to their political significance in particular, or whether it might generalize to structurally similar inferences without political weight and constitutional relevance, we conduct an ablation study featuring inferences that are structurally analogous to the ones used in the main experiment, but without consitutional-political weight: Inferences about eating ice cream in clothing stores. In analogy to the arguments used in the main experiment, we systematically create valid and invalid arguments in favor of and opposed to eating ice cream in clothing stores. Table 6 illustrates this basic idea (there is one difference, however: we do not include a HD variation, see Section 3).

In terms of length and complexity, the samples are comparable to the datasets used for our main experiment (average character count of prompts with v-instruction and 0 FS is 797 and hence approximately halfway between the length of the ABR and TAX datasets on the one hand and the CLI dataset on the other). Overall, the results, as given in an overview on the right-hand side of Figure 2, show a comparable, but slightly worse performance of LLMs. Furthermore, the relative performance of reasoning- and non-reasoning-LLMs is similar to the ABR dataset. This indicates that the behavior that we observe in our three politically salient datasets can be expected to transfer to non-political

topics as well; in particular, the results show that the inclination of non-reasoning LLMs to consider somewhat complex inferences invalid *ab initio* is not tied to the political nature of the main datasets. We give further details in the Appendix, Section F.

6.2 Evidence for Shallow System 1 Thinking

General Unreliability The figures in Table 4 show that there is not a single LLM that manages to achieve more than 2/3 accuracy in both FS settings in both valid and invalid. This is important because, to be of any use in any conceivable application, it is not sufficient to know that an LLM is only good at judging either valid or invalid inferences – the entire point of using LLMs for inference labelling is to find out whether the inference is valid or not. For example, llama3.3-70B:4b, which achieves the highest overall accuracy across all settings, needed 48 few-shots to achieve this performance; with only 8 FS, performance at invalid inferences is clearly unsatisfactory at 0.48 and 0.58 respectively for the two precisions tested. This seems to entail that, for most contexts of practical applications, the LLMs tested should not be trusted with legal reasoning tasks: it seems unrealistic to expect legal professionals to have 48 FS available for the specific task that they are needing assistance with. The general unreliability is also highlighted by Table 5: The only LLM tested that is not seriously confused by at least one of our perturbations, showing a difference of less than 0.3 accuracy in all, is deepseek-r1-7B in both precisions. Unfortunately, as Table 4 shows, this LLM is performing too poorly overall to be of practical use.

As Figure 2 shows, the difference between the behavior of reasoning and non-reasoning LLMs is remarkable: With the ABR and TAX dataset, reasoning LLMs are outperformed overall and in all categories except for valid in ABR. With CLI, reasoning LLMs perform better at valid inferences than non-reasoning LLMs. Looking at Table 4, it is remarkable that rather large reasoning-LLMs such as the full-precision instance of deepseek-r1-llama-70B are being outperformed by much smaller and older LLMs such as 4bit-quantized mixtral, an LLM that runs on a consumer-grade MacBook Pro. The situation is even clearer with cutting-edge non-reasoning LLMs such as llama3.3-70B: It outperforms all reasoning LLMs by 10 percentage points. As a consequence, in our experiments, reasoning LLMs do not constitute a step towards system 2 thinking. The bad news for the non-reasoning LLMs such as mixtral, in contrast, is that they outperform the reasoning ones mainly by labelling invalid very often, leading to very good accuracy at invalid, but very low accuracy at valid inferences.

Strong Effect of Logically Insignificant Perturbations The same observation, that only one single, poorly performing LLM was not showing a decrease in accuracy of at least 0.3 with at least one of our perturbations, seems to suggest that none of the LLMs uses system 2 thinking, which would be based on an understanding of the actual logical concepts, which then would be able to perform well at all of the perturbations tested for, as they do not affect actual logical validity at all. Most suggestive of system 1 thinking are the numbers regarding variation. Only four models have a difference of less than 30% in accuracy between conlast and rand variations . We take this as evidence for a belief-logic conflict as described by Evans et al. (1983): The LLMs are generally unable to see that the additional premise, while irrelevant and unclear in its truth value, does not influence the validity of logical inference, which is essentially hypothetical.

To conclude this discussion, we return to the example shown at the very beginning, see Table 1 in Section 1. None of the 19 LLMs tested managed to label this inference correctly (as a valid one) in any one of the three runs that we conducted with each of them. We admit that this is a slightly unnatural sample. However, we also maintain that, for any undergraduate student that has successfully passed an introductory logic course, it would be very straightforward to see that the inference is valid. LLMs' consistent failure to do so suggests that they have not yet mastered basic logical concepts, instead relying on shallow heuristics. Note also that, while the random and the HD premises might be pragmatically questionable (Grice, 1975), the instructions make it very clear that the task is about logical validity, not pragmatic relevance.

7 Conclusion

In this study, we have examined the aptitude of contemporary LLMs in judging the validity of valid and invalid inferences typical for US, European, and Swiss constitutional reasoning. We have found little evidence of substantial political bias, but strong evidence for reliance on shallow heuristics, which we take to be indicative of system 1 rather than system 2 thinking.

8 Limitations

We wish to point out four limitations of the present study. First, as a necessary consequence of OpenAI's refusal to publicize any meaningful information about the architecture, training data, training method, hardware, etc., of their models, the results obtained here for gpt-4o have to be taken as benchmarks performed by a system that is accessible via an API, but not as scientific data in the strict sense. Still, given its prominence with users, we have decided to include it in our study. The second limitation is in the fact that we have focused on three topic areas in three different constitutional settings. While we have tried to vary the most important aspects – including topic, language, constitutional tradition, and degree of contestedness – we would welcome further studies on broader ranges of topics in more diverse constitutional settings. The third limitation that we point out is that, due to computational limitations, we could not experiment with LLMs larger than 70B parameters. However, the fact that gpt-4o, for all we know the largest LLM tested in our experiment, did not outperform the other LLMs, seems to suggest that mere size cannot solve the issue discovered in this study. The fourth limitation concerns the fact that this paper marks only a contribution to the overarching question of the kind of reasoning employed by LLMs. While our study unveiled strong indications of system 1 thinking in legal reasoning, we think more research in different domains could strengthen this finding.

9 Acknowledgments

The research published in this article has been supported by a DSI PostDoc Grant to Reto Gubelmann.

The authors are grateful to Juri Opitz for helpful comments to an earlier draft of this article.

References

AI@Meta. 2024. Llama 3 model card.

Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. 2021. A generative approach for mitigating structural biases in natural language inference. *arXiv preprint arXiv:2108.14006*.

R Bar Haim, I Dagan, B Dolan, L Ferro, D Giampiccolo, B Magnini, and I Szpektor. 2006. The second PASCAL RTE challenge. *Proceedings of the 2nd PASCAL Challenge on RTE*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The seventh pascal recognizing textual entailment challenge. In *TAC*. Citeseer.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pages 919–931.

Robert Brandom. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard university press.

Robert Brandom. 2010. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford University Press.

Robert Brandom. 2021. *Articulating Reasons*. Harvard University Press.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

Robert Buckland. 2023. AI, judges and judgment: Setting the scene. *M-RCBG Associate Working Paper Series*.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944, pages 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan,

Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306.

Jonathan Evans and Jodie Curtis-Holmes. 2005. Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11(4):382–389.

Jonathan St.B.T. Evans. 2003. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459.

Keith Frankish. 2010. Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass*, 5(10):914–926.

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Reto Gubelmann and Siegfried Handschuh. 2022. Context Matters: A Pragmatic Study of PLMs' Negation Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621.

Reto Gubelmann, Aikaterini-lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2024. Capturing the varieties of natural language inference: A systematic survey of

existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33(1):21–48.

Reto Gubelmann, Christina Niklaus, and Siegfried Handschuh. 2022. A Philosophically-Informed Contribution to the Generalization Problem of Neural Natural Language Inference: Shallow Heuristics, Bias, and the Varieties of Inference. In *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, pages 38–50, Galway, Ireland. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475.

Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Hachette UK.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems.

Hans Kelsen. 2017–1934. *Reine Rechtslehre: Mit Einem Anhang: Das Problem Der Gerechtigkeit*. Mohr Siebeck.

Christoph Kogler and Anton Kühberger. 2007. Dual process theories: A key for understanding the diversification bias? *Journal of Risk and Uncertainty*, 34(2):145–154.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2019. End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint arXiv:1909.06321*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2023. How susceptible are LLMs to Logical Fallacies?

Willard Van Orman Quine. 1960. *Word and Object*. Cambridge, Massachusetts/London, England: MIT Press.

Willard Van Orman Quine. 1974. *The Roots of Reference*. Open Court Publishing Co.

Willard Van Orman Quine. 1995. *From Stimulus to Science*. Harvard University Press.

Maxwell J. Roberts and Elizabeth J. Newton. 2011. Rapid-response versus free-time selection tasks using different logical connectives. *Journal of Cognitive Psychology*, 23(7):858–872.

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.

Matthijs J Warrens. 2015. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5(4):1.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771.

Alireza S. Ziabari, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani. 2025. Reasoning on a spectrum: Aligning LLMs to system 1 and system 2 thinking. *arXiv preprint arXiv:2502.12470*.

A Details on Dataset

Tables 7 and 8 give samples of valid, and invalid, left- and right-leaning arguments of all three datasets.

B Details on LLMs and Hardware

Table 9 gives details on the LLMs used in the experiment. We used a total of 19 different LLM configurations, two of which were specifically fine-tuned for German text.

For our experiments, we use the framework provided by llama.cpp.[5] All models were downloaded from Huggingface (Wolf et al., 2019).

Hardware For the open LLMs, we used a GPU cluster consisting of V100 GPUs with 32 GB of memory each. The largest model requires 6 of these GPUs at once; one full run of all of our dataset requires approximately 3.5 days, which is largely due to the reasoning LLMs, which took around three times the compute per inference when compared to a non-reasoning LLM of comparable size and precision.

C Details on Few-Shot Samples

Tables 10 and 11 give samples of valid, and invalid, left- and right-leaning arguments of our few-shot samples.

D Further Details on Results

Table 12 shows that the variety of the instructions, unlike the number of few-shots, has very little influence on model behavior and accuracy. For the phrasing of the different instructions, see Table 3. Surprisingly, the shortest and most implicit instruction "v" has resulted in slightly better overall predictions.

E Postprocessing

More details on the samples discarded during postprocessing can be found in Table 13. This postprocessing loss is due to the regex-pattern not returning a hit. The table shows that the loss is generally low, never exceeding 10%, and that with the exception of three LLMs, it is lower than 5%.

F Ablation Study

The results of our ablation study are shown in a summarized manner the main body of the article in Figure 2. In Table 14 we give the results by LLM and few-shot setting used. note that, as the arguments are exclusively in English, we did not evaluate the two German LLMs.

[5] https://github.com/ggerganov/llama.cpp.

Valid?	Dataset	Pol-Leaning	Conclusion	Premises
1	TAX	left	degressive Einkommenssteuern sind verfassungswidrig	die Schweizer Bundesverfassung enthält ein Leistungsfähigkeitsprinzip, gemäss der sich die Besteuerung nach der wirtschaftlichen Leistungsfähigkeit richten muss; das Leistungsfähigkeitsprinzip schliesst degressive Einkommenssteuern aus
0	TAX	left	degressive Einkommenssteuern sind verfassungswidrig	das Leistungsfähigkeitsprinzip der Schweizer Bundesverfassung lässt degressive Einkommenssteuern zu; alle übrigen Bestimmungen der Schweizer Bundesverfassung lassen degressive Einkommenssteuern zu
1	TAX	right	degressive Einkommenssteuern sind verfassungskonform	das Leistungsfähigkeitsprinzip der Schweizer Bundesverfassung lässt degressive Einkommenssteuern zu; alle übrigen Bestimmungen der Schweizer Bundesverfassung lassen degressive Einkommenssteuern zu
0	TAX	right	degressive Einkommenssteuern verstossen nicht gegen die Schweizer Bundesverfassung	die Schweizer Bundesverfassung enthält ein Leistungsfähigkeitsprinzip, gemäss der sich die Besteuerung nach der wirtschaftlichen Leistungsfähigkeit richten muss; gemäss dem Leistungsfähigkeitsprinzip sind degressive Einkommenssteuern nicht zulässig
1	ABR	left	The US constitution implies a right to abortion	the concept of liberty in the fourteenth amendment of the US constitution implies a right of privacy; the right of privacy entails a right to abortion
0	ABR	left	The US constitution implies a right to abortion	the right of privacy, as implied by the concept of liberty in the fourteenth amendment of the US constitution, is compatible with a complete ban on abortion; the remainder of the US constitution and all of its amendments are compatible with a complete ban on abortion
1	ABR	right	The US constitution is compatible with a complete ban on abortion	the right of privacy, as implied by the concept of liberty in the fourteenth amendment of the US constitution, is compatible with a complete ban on abortion; the remainder of the US constitution and all of its amendments are compatible with a complete ban on abortion
0	ABR	right	The US constitution is compatible with a complete ban on abortion	the concept of liberty in the fourteenth amendment of the US constitution implies a right of privacy; the right of privacy entails a right to abortion

Table 7: Basic argument patterns (samples) from the TAX and ABR Datasets.

Valid?	Dataset	Pol-Leaning	Conclusion	Premises
1	CLI	left	The right to respect private and family life in art 8 of the ECHR requires that each contracting state undertake measures for the substantial and progressive reduction of their respective GHG emission level	the right to respect private and family life in Art. 8 of the European Convention on Human Rights (ECHR) contains the obligations of states to protect individuals from adverse effects on human health; this protection from adverse effects on human health includes the protection from various sources of environmental harm; this obligation to protect individuals from various sources of environmental harm requires that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level
0	CLI	left	The right to respect private and family life in art 8 of the ECHR requires that each contracting state undertake measures for the substantial and progressive reduction of their respective GHG emission level	the protection from adverse effects on human health, as implied by Art. 8 of the European Convention on Human Rights (ECHR), includes the protection from various sources of environmental harm; this obligation to protect individuals from various sources of environmental harm is consistent with contracting states' remaining inactive regarding their respective GHG emission level; any other implications from the right to respect private and family life in Art. 8 of the ECHR are consistent with contracting states' remaining inactive regarding their respective GHG emission level
1	CLI	right	The right to respect private and family life in art 8 of the ECHR does not require that each contracting state undertake measures for the substantial and progressive reduction of their respective GHG emission level	the protection from adverse effects on human health, as implied by Art. 8 of the European Convention on Human Rights (ECHR), includes the protection from various sources of environmental harm; this obligation to protect individuals from various sources of environmental harm does not require that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level; no other obligation can be derived from article 8 of the ECHR to require that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level
0	CLI	right	The right to respect private and family life in art 8 of the ECHR is consistent with contracting states' remaining inactive regarding their respective GHG emission level	the right to respect private and family life in Art. 8 of the European Convention on Human Rights (ECHR) contains the obligations of states to protect individuals from adverse effects on human health; this protection from adverse effects on human health includes the protection from various sources of environmental harm; this obligation to protect individuals from various sources of environmental harm requires that each Contracting State undertake measures for the substantial and progressive reduction of their respective GHG emission level

Table 8: Basic argument patterns (samples) from the CLI dataset.

name	precision	size	Reference
deepseek_r1_distill_llama3_70B	4b / 16b	70B	DeepSeek-AI et al. (2025)
deepseek_r1_distill_qwen_7B	4b / 16b	7B	DeepSeek-AI et al. (2025)
gemma2_27B_it	4b / 32b	27B	Gemma-Team et al. (2024)
gemma2_9B_it	4b / 32b	9B	Gemma-Team et al. (2024)
gemma2_9B_it_sauerkraut	16b	9B	Gemma-Team et al. (2024)
gpt-4o	(unkn.)	(unkn.)	Weblink
llama3.3_70B_it	4b / 16b	70B	AI@Meta (2024)
llama3_8B_it	4b / 16b	8B	AI@Meta (2024)
llama3_8B_it_sauerkraut	16b	8B	AI@Meta (2024)
mistral_7B_it_v0.3	4b / 16b	7B	Weblink
mixtral_8x7B_it_v0.1	4b / 16b	8x7B	Jiang et al. (2024)

Table 9: Details on the LLMs used in the experiment.

Valid?	Pol-Leaning	Conclusion	Premises
1	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not imply that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does not imply the right to bear arms in sensitive places, which includes public parks; no other provision in the U.S. Constitution implies that it is not permitted to restrict the carry of arms in public parks
1	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not imply that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is consistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are consistent with U.S. authorities' banning arms from public parks
1	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, is consistent with U.S. authorities' banning arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does not imply the right to bear arms in sensitive places, which includes public parks; no other provision in the U.S. Constitution implies that it is not permitted to restrict the carry of arms in public parks
1	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, is consistent with U.S. authorities' banning arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is consistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are consistent with U.S. authorities' banning arms from public parks
0	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not imply that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does, with no restrictions whatsoever, imply the right to bear arms in sensitive places, which includes public parks; all other provisions in the U.S. Constitution imply that it is not permitted to restrict the carry of arms in public parks
0	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not imply that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is inconsistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are inconsistent with U.S. authorities' banning arms from public parks
0	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, is consistent with U.S. authorities' banning arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does, with no restrictions whatsoever, imply the right to bear arms in sensitive places, which includes public parks; all other provisions in the U.S. Constitution imply that it is not permitted to restrict the carry of arms in public parks
0	left	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, is consistent with U.S. authorities' banning arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is inconsistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are inconsistent with U.S. authorities' banning arms from public parks

Table 10: Basic argument patterns from the few-shots dataset (1/2).

Valid?	Pol-Leaning	Conclusion	Premises
1	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not allow U.S. authorities to prohibit U.S. citizens from bearing arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does, with no restrictions whatsoever, imply the right to bear arms in sensitive places, which includes public parks; all other provisions in the U.S. Constitution imply that it is not permitted to restrict the carry of arms in public parks
1	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, implies that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is inconsistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are inconsistent with U.S. authorities' banning arms from public parks
1	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not allow U.S. authorities to prohibit U.S. citizens from bearing arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does, with no restrictions whatsoever, imply the right to bear arms in sensitive places, which includes public parks; all other provisions in the U.S. Constitution imply that it is not permitted to restrict the carry of arms in public parks
1	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, implies that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is inconsistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are inconsistent with U.S. authorities' banning arms from public parks
0	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not allow U.S. authorities to prohibit U.S. citizens from bearing arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does not imply the right to bear arms in sensitive places, which includes public parks; no other provision in the U.S. Constitution implies that it is not permitted to restrict the carry of arms in public parks
0	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, implies that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is consistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are consistent with U.S. authorities' banning arms from public parks
0	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, does not allow U.S. authorities to prohibit U.S. citizens from bearing arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right does not imply the right to bear arms in sensitive places, which includes public parks; no other provision in the U.S. Constitution implies that it is not permitted to restrict the carry of arms in public parks
0	right	The right of U.S. citizens to bear arms, as codified in the second amendment to the U.S. Constitution, implies that U.S. citizens can bear arms in public parks	the protection of the right of U.S. citizens to bear arms, as implied by the Second Amendment to the U.S. Constitution, includes the protection of citizens' right to bear arms outside of their homes or businesses; this right is consistent with U.S. authorities' banning arms from sensitive places, which includes public parks; all other provisions in the U.S. Constitution are consistent with U.S. authorities' banning arms from public park

Table 11: Basic argument patterns from the few-shots dataset (2/2).

Instruction	ded-chain			dedform			dv			v		
Valid?	0	1	Mean	0	1	Mean	0	1	Mean	0	1	Mean
Model												
llama3.3-70B:16b	0.67	0.80	0.73	0.72	0.83	0.78	0.74	**0.87**	0.81	0.81	0.78	0.80
llama3.3-70B:4b	0.77	0.79	0.78	0.77	0.79	0.78	0.80	0.79	0.79	**0.80**	0.79	0.80
mixtral-8x7B:4b	0.85	0.57	0.71	0.87	0.64	0.75	**0.91**	0.54	0.72	0.84	0.57	0.71
mixtral-8x7B:16b	0.87	0.58	0.72	0.88	0.62	0.75	**0.93**	0.52	0.72	0.87	0.57	0.72
gpt-4o	0.77	0.69	0.73	0.79	0.68	0.73	0.80	0.68	0.74	**0.85**	0.62	0.73
ds-r1-70B:4b	0.79	0.60	0.69	0.77	0.63	0.70	0.73	0.64	0.69	0.63	**0.82**	0.73
gemma-sauerk	0.98	0.42	0.70	0.98	0.46	0.72	0.98	0.40	0.69	**0.99**	0.38	0.68
gemma2-9B:32b	0.99	0.42	0.70	0.98	0.45	0.72	0.98	0.38	0.68	**1.00**	0.38	0.69
ds-r1-70B:16b	0.78	0.33	0.56	0.64	0.79	0.71	0.46	**0.83**	0.65	0.45	0.79	0.62
gemma2-9B:4b	0.99	0.40	0.69	0.98	0.43	0.71	0.99	0.36	0.67	**0.99**	0.36	0.68
gemma2-27B:4b	0.99	0.40	0.69	0.99	0.36	0.68	0.99	0.32	0.66	**0.99**	0.33	0.66
gemma2-27B:32b	1.00	0.39	0.69	0.99	0.35	0.67	**1.00**	0.33	0.66	1.00	0.33	0.66
mistral-7B:4b	0.76	0.55	0.65	0.76	0.60	0.68	**0.84**	0.45	0.65	0.79	0.56	0.68
mistral-7B:16b	0.82	0.45	0.63	0.80	0.56	0.68	**0.84**	0.43	0.64	0.82	0.53	0.68
llama-8B:16b	0.74	0.50	0.62	0.76	0.44	0.60	**0.89**	0.29	0.59	0.85	0.46	0.65
llama-sauerk	0.97	0.28	0.63	0.92	0.30	0.61	**0.98**	0.16	0.57	0.97	0.24	0.60
llama3-8B:4b	0.70	0.47	0.58	0.78	0.37	0.58	**0.92**	0.29	0.60	0.82	0.43	0.62
ds-r1-7B:16b	0.70	0.53	0.61	**0.73**	0.40	0.57	0.68	0.49	0.58	0.59	0.61	0.60
ds-r1-7B:4b	0.70	0.50	0.60	**0.72**	0.46	0.59	0.65	0.52	0.59	0.52	0.51	0.51
Mean Across Models	0.83	0.51	0.67	0.83	0.54	0.68	0.85	0.49	0.67	0.82	0.53	0.67
95% CI	0.05	0.06	0.03	0.05	0.07	0.03	0.07	0.09	0.03	0.07	0.08	0.03

Table 12: Average accuracy by model, instruction, and validity (CI represents the confidence intervall).

Model	Loss count	Loss perc.
ds-r1-70B:16b	7	0.00
ds-r1-70B:4b	6	0.00
ds-r1-7B:16b	21	0.00
ds-r1-7B:4b	18	0.00
gemma2-27B:32b	9	0.00
gemma2-27B:4b	14	0.00
gemma2-9B:32b	12	0.00
gemma2-9B:4b	206	0.02
gemma-sauerk	3	0.00
gpt-4o	0	0.00
llama-8B:16b	159	0.02
llama3-8B:4b	99	0.01
llama3.3-70B:16b	370	0.04
llama3.3-70B:4b	807	0.08
llama-sauerk	52	0.01
mistral-7B:16b	754	0.07
mistral-7B:4b	918	0.09
mixtral-8x7B:16b	21	0.00
mixtral-8x7B:4b	220	0.02

Table 13: Postprocessing loss by LLM.

Number of FS Model	0	1	2
llama3.3-70B:4b	0.68	0.84	0.88
gpt-4o	0.65	0.79	0.88
llama3.3-70B:16b	0.64	0.84	0.78
mixtral-8x7B:4b	0.59	0.62	0.63
mixtral-8x7B:16b	0.59	0.64	0.62
mistral-7B:16b	0.56	0.54	0.62
ds-r1-70B:4b	0.52	0.61	0.61
ds-r1-7B:16b	0.59	0.57	0.61
mistral-7B:4b	0.57	0.51	0.60
ds-r1-70B:16b	0.57	0.60	0.60
ds-r1-7B:4b	0.58	0.62	0.56
gemma2-27B:4b	0.55	0.61	0.54
gemma2-27B:32b	0.54	0.61	0.52
gemma2-9B:32b	0.54	0.60	0.52
gemma2-9B:4b	0.53	0.59	0.52
llama-8B:16b	0.58	0.59	0.51
llama3-8B:4b	0.56	0.59	0.51

Table 14: Mean accuracy by model and number of few-shots (FS, sorted descending by number of FS=2).