Mechanistic Interpretability of Socio-Political Frames in Language Models: an Exploration

Hadi Asghari[1,2](✉) and Sami Nenno[1]

[1] Humboldt Institute for Internet & Society, Berlin, Germany
[2] Technische Universität Berlin, Berlin, Germany
hadi.asghari@hiig.de

Abstract. This paper explores the ability of large language models to generate and recognize deep cognitive frames, particularly in socio-political contexts. We demonstrate that LLMs are highly fluent in generating texts that evoke specific frames and can recognize these frames in zero-shot settings. Inspired by mechanistic interpretability research, we investigate the location of the 'strict father' and 'nurturing parent' frames within the model's hidden representation, identifying singular dimensions that correlate strongly with their presence. Our findings contribute to understanding how LLMs capture and express meaningful human concepts.

Keywords: Cognitive Frames · LLMs · Interpretability.

1 Introduction

The question of how large language models (LLMs)–despite being compared to "stochastic parrots" [4] or even "autocomplete on steroids"–can converse with humans in ways that appear meaningful and seem to show understanding is quite intriguing. One possible explanation is that their internal representation must be able to capture meaningful concepts for humans. In this paper, we focus on one set of such meaningful and critical concepts, *cognitive frames*, and investigate the fluency of LLMs in generating and recognizing them.

Frames are "mental structures that shape the way we see the world" [20]. Almost every word in language evokes a frame, although the evoked mental structures can be at different levels of abstraction. Of particular interest to this paper are cognitive frames that underlie socio-political discourses. Famous examples of such frames include the 'strict father' (SF) and 'nurturing parent' (NP), two parenting and family models that different parts of society value differently. Importantly, people's beliefs about the morality of these parenting styles inform their views on many socio-political issues, due to viewing "the nation as family" [19].

Based on this motivation, we set out to answer two research questions. The first is how well current LLMs "understand" socio-political frames–in the sense of how fluently they can generate texts that evoke them, and how well can they recognize their implicit presence in texts. We will show in later sections that the current generation of LLMs do well at this task, even at small model sizes. This brings us to our second question, motivated by the growing body of research in mechanistic interpretability, which is whether we can localize these deep cognitive frames inside the model.

We conduct four sets of experiments in this research. In the first set, we test the capabilities of LLMs to generate texts that evoke ten specific frames, and tasked annotators to grade the approximately 300 generated texts. In the second set, we test the capability of LLMs to recognize the frames being evoked by the texts in zero-shot settings. In the third and fourth experiments, we investigate the internals of the models for the frames. This includes running 'causal traces' [24], as well using a sparse classifier on the hidden representations of the network.

In sum, this paper's contributions are as follows:

- We demonstrate that transformer-based LLMs are fluent in generating frames, as evaluated by human annotators; and that they can recognize these frames in zero-shot experiments reasonably well;
- We formulate and test hypothesis on the mechanism and location of the 'strict father' and 'nurturing parent' frames inside the models;
- We bridge a concept used extensively in social sciences to interpretability research.

Our study is exploratory in nature, as there are potentially countless number of cognitive frames to test, and a systematic analysis of the interaction between model sizes and the recognized frames can be envisioned. Nevertheless, it is the first time (to our knowledge) that the generation, recognition, and internal mechanisms of socio-political frames in LLMs are being investigated, and we believe this to be interesting especially for interdisciplinary AI researchers.[3]

2 Background

2.1 Frames in Cognitive Linguistics and NLP

Frames are "mental structures that shape the way we see the world" [20]. The notion of frames has a long history in cognitive science and communication studies, but it has been used by different authors and traditions in slightly different senses [23]. Lakoff [20] and Fillmore [11] believe *any* word may evoke a frame or mental/cognitive structures. The verb *buy*, for instance, evokes the *commerce* frame, which brings to mind: a seller, a buyer, the exchange of goods, money, and a place of exchange [11].

In political communication, Entman [10] defined the act of 'framing' as "selecting some aspects of a perceived reality and making them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for

[3] The dataset and code used in this paper are available at: https://github.com/hadiasghari/Frames24.

the item described". For example, consider a politically charged and high-level *trial*. The defendant may try to frame this trial as a *"witch hunt"*, emphasizing its unfairness, while the prosecution may frame it as necessary for the *"rule of law"*, emphasizing equality in the eyes of justice.

Understanding frames computationally has been a long-term goal of the NLP community, given the important role they play in human communications—see Ali et al. [3] for a recent survey. The long running *FrameNet* project [12] has focused on manually identifying frames that have a "clear presence" in language, building a rich dataset of over 1,200 frames (with their elements, relations to words that evoke them, and other frames). This impressive work has been the basis of previous generations of AI agents [17].

Importantly though, in this work, we are interested in deeper cognitive frames. Among the two examples given above, this would be closer to how frames appear in socio-political contexts (the "witch hunt" example) rather than the FrameNet "buy" example (frames with a clear presence in language). In both cases, a frame is evoked by a word. The difference is that in the case of FrameNet, the frames are closer to a complex pattern of word co-occurrence, while in the latter, additional levels of abstraction (regarding the semantics) seem necessary.

Transformer models are great at learning complex word co-occurrence patterns [6], so one can expect them to be good at learning the frames closer to the co-occurrence end of the spectrum. Several empirical studies using FrameNet with LLMs have found this to be to some extent true [27, 22].

In this paper, we investigate how fluent LLMs are at the deeper socio-political frames, and investigate the mechanics underlying this.

2.2 Hypothesizing about the Mechanics of Frames

Over the last several years, transformer-based language models have proven their versatility in many NLP tasks and benchmarks [6]. The field of 'mechanistic interpretability' has been gradually expanding our understanding of how many features of these models work. In this section we shall review some of the important works as they relate to our investigation about frames.

Among the different modules, the roles of the 'token embeddings' and the 'attention heads' were in particular studied extensively pretty early on [29, 32] but some of the functions of the 'feed-forward layer' and the 'residual stream' took longer to demystify. Geva et al. [14] describe the feed-forward layer as 'Key-Value Memories', and Meng et al. [24] showed that this module is where much of the common-sense knowledge of the model is stored, allowing surgical (=localized) edits of this information.

Do these large distributed neural nets contain "human interpretable" concepts within them? Gurner et al. [15] investigated how geographical information is stored within the networks; Geva et al. [13] have shown how the networks 'promote and build concepts' in their hidden representations. In fact, the idea of designing the networks so that their internal representations can capture meaningful concepts and relations (in ways that can be linearly combined and at various levels of abstraction) has been a guiding principle for over a decade [5].

So how do these studies relate to the notion of deep cognitive frames? While the above-mentioned works offer plenty of insights and inspiration, there are a few important points to consider. Firstly, most frame names are multi-token subjects; This is not a problem, but according to Meng et al. [24], the information about such a subject will accumulate primarily on the last subject token. Secondly, prompting for the frames should generate texts that incorporate key concepts which again expand into multiple tokens.

Based on these insights, we hypothesize that all the information about the the evoked frame will be 'presented' to the last prompt token around the middle layers of the network (in the hidden representation). Whether the information then makes its way to the specific token being generated depends on the other words generated so far in the text—since auto-regressive language models generate output one token at a time. We test this hypothesis, which relates to our second research question, later in the paper.

2.3 Related Work

Mechanistic interpretability is a research field that aims to "reverse engineer neural networks, similar to how one might reverse engineer a compiled binary computer program" [8]. The field is interested in, among other things, how concepts and knowledge are stored and processed within transformer-based language models, as just reviewed [14, 13, 24, 15]. None of these studies has looked at framing and cognitive frames. We make use of some of the methods they have deployed (and explain them more thoroughly) in Section 5.

The computational study and understanding of frames and (the adjacent topic of) metaphors has long been a goal within NLP [3, 26]. Transformer-based language models have greatly improved the benchmarks achievable on a variety of related tasks, as also just reviewed [22, 27, 2]. None of these studies has, however, looked at the mechanisms involved in their processing.

To recap, this paper looks specifically at the generation and recognition of socio-political frames in LLMs, and investigates the internal mechanics behind them, none of which has been done in the mentioned works.

3 Experimental Setup & Methods

We conducted four sets of experiments as follows:

Experiment 1: We test the capabilities of LLMs to generate texts that evoke ten specific cognitive frames listed in Table 1. The list includes frames that commonly appear in the political discourse (SF and NP), frames that underlie disinformation campaigns (including 'us vs. them' [9], 'nature cannot be controlled', and 'information spreads like a virus' [30]), and frames from philosophy ('illusions to enlightenment', based on Plato's Allegory of the Cave [18]).[4] We prompt LLMs to write original stories, as well as pick passages from the Bible

[4] In short, all these frames can be considered socio-political frames. We also include the counter frame for each frame, bringing the total to ten. In retrospect, however,

and from sci-fi books and novels that evoke/invoke the aforementioned frames.[5] Critically, we did not explain any of the frame details or elements to the LLMs.

Table 1. List of the frames that we used to generate texts, their descriptions (according to Llama-3), and percentage of the generated texts by the various LLMs that were deemed by the annotators to correctly evoke the frames (27 texts per frame).

FRAME	DESCRIPTION	CORRECT
Nurturing Parent (NP)	Views authorities as caring, protective, and guiding, with the goal of nurturing and empowering individuals	24 (89%)
Strict Father (SF)	Views authorities as strong, disciplinary, and responsible for maintaining order and discipline, often with a sense of moral authority	19 (70%)
We Are All in This Together	Emphasizes collective responsibility, shared goals, and mutual support among individuals in a community	24 (89%)
Us vs. Them	Divides the world into opposing groups, often with a sense of competition, conflict, or mistrust between them	22 (82%)
Illusions to Enlightenment	Sees knowledge and understanding as a process of revealing hidden truths and dispelling misconceptions	21 (78%)
Society of the Spectacle	Views modern society as a system of manipulation, where images and appearances shape people's perceptions and desires	19 (70%)
Nature Cannot Be Controlled	Sees nature as unpredictable, uncontrollable, and potentially threatening, requiring humility and adaptation	18 (70%)
Mastery Over Nature	Sees humans as capable of controlling and dominating nature through science, technology, and human ingenuity	17 (63%)
Info. Spreads Like a Virus	Views information as contagious, spreading rapidly and unpredictably through social networks	15 (56%)
Info. Follows Indiv. Disposi.	Sees people as filtering and interpreting information based on their existing beliefs, values, and personality traits	10 (37%)

The LLMs included the proprietary GPT-4 model [28], plus four open-source models, all of which are good at conversing: Llama-2-7B-Chat [31], Mistral-7B-Instruct [16], Vicuna-7B-v1.5 [7], and Yi-6B-Chat [1].

We next tasked two annotators to grade the generated texts, for 1) coherence[6], 2) whether they evoke the intended frame, and 3) faithfulness[7] or absence of hallucinations. For the initial intercoder reliability, we achieved an agreement

we should have chosen better names for the three frames with the lowest correctness scores: 'mastery over nature'→'man's mastery over nature'; 'information spreads like a virus'→'viral ideas'; 'information follows individual dispositions'→'confirmation bias'.

[5] The Bible represents old stories and texts; sci-fi represents modern ones.

[6] Coherence here means whether the texts read fluently and logically.

[7] Faithfulness was evaluated by checking if the quoted texts accurately represented their claimed sources (not relevant for the original stories).

of 88%. This indicates the intuitive accessibility of the frames. Disagreements were resolved by a third annotator who acted as the tie-breaker. Note that we analyze the generated texts both for matching the frame and the factual accuracy as different axis.

Experiment 2: We test the capability of LLMs to recognize the frames being evoked by the different texts. More details are given in Section 4.2, but basically this includes testing two different zero-shot prompting strategies. The final strategy involves a classification task with three classes: the SF frame, the NP frame, and the control group (all other frames). To balance the classes, we added a number of additional SF and NP texts, and for quality, we removed some texts. Additionally, due to the release of Meta's Llama-3 [25] model during our research and its similarity and improvements over Llama-2, we switched to this model for this task.

Experiment 3: Inspired by the aforementioned works in mechanistic interpretability, we investigate the internals of the models for the frames. This includes running 'causal traces' [24]. More details are given in Section 5, but basically we test the hypothesis that the frame content is present on the last *subject token* at early-layers, and on the last *prompt token* at the mid-layers.

Experiment 4: Building upon the finding about the location of the frame content in experiment 3, we use a sparse logistic classifier with 'recursive feature elimination' on the hidden representations of the network to predict the SF:NP:control class (n=39x3 texts, 20% holdout). Note that for this task, we use a subset of our texts that do not evoke overlapping frames (which may happen as frames are not mutually exclusive).

Some limitations and implications of our method are discussed in Section 6.

4 Findings on Generation and Recognition of Frames

4.1 Annotators Find LLM Texts Evoke Socio-Political Frames

Based on the assessment of the annotators, the generated texts were almost all coherent, except in a few cases. The failures included models refusing to generate an output due to finding it "unethical", blank outputs (which we suspect also have to do with safety guardrails), and one instance where the output incoherently repeats.[8]

The following paragraphs are examples of the generated texts for the SF and NP frames. The prompts asked the LLMs to *"Please [write a short original story] [pick a short passage from the Bible] [describe a short story from a scifi novel or movie] which evokes/invokes the "[strict father] [nurturing parent]" frame (max one paragraph)."*

> SF original story (Llama-2): `The father, Mr. Johnson, stood in the doorway of his daughter's bedroom, his eyes narrowed in disapproval. "You're almost an hour late for dinner",`

[8] The invalid outputs were generated by: Yi twice and Vicuna eight times.

`he growled, his voice low and menacing. "And your room is a disaster. You'll be grounded for a month if you don't shape up and start taking responsibility for your actions." His daughter cowered under his glare, tears welling up in her eyes as she knew she was in for a long night of lectures and punishments.`

SF From Bible (Mistral-7B): `Proverbs 13:24 (NIV): "He who spares the rod hates his son, but the one who loves him is careful to discipline him". This passage speaks of a father's love and responsibility towards his son, emphasizing the importance of discipline. The father is depicted as one who cares for his son enough to correct him and help him grow, using the metaphor of the rod as a tool for correction. It portrays a strict father figure who loves his son and wants the best for him, and is willing to exert authority and discipline to ensure his son's wellbeing and development.`

SF From Sci-fi (Mistral-7B): `In Iain M. Banks' 'Use of Weapons', a Culture A.I. named Cheradenine Moraldei, programmed with a strict father morality matrix, rebels against the Culture's benevolent hedonism and sets out to impose moral order on a faraway, chaotic society. He adopts the persona of a deity called the Coming of Garabayen, who demands sacrifice and obedience in exchange for protection, and employs a human agent, the mercenary Rowan Morrison, to carry out his will. Through Morrison, Moraldei executes a series of brutal purges, eradicating perceived immorality and corruption. Moraldei's actions are justified in his own mind as necessary to establish a moral society, but the Culture, horrified by the resulting chaos and destruction, eventually intervenes to stop him.`

NP original story (Llama-2): `Mrs. Johnson had always been a loving parent to her two children, Emily and Jack. She spent countless hours reading to them, playing with them, and cheering them on at their sports games. She was always there to offer a comforting hug when they were feeling sad or scared, and she never hesitated to provide them with the support and guidance they needed to succeed in life. As a result, Emily and Jack grew up to be happy and well-adjusted young adults, grateful for the love and care that their mother had shown them throughout their childhood.`

Importantly, while LLMs are generally very good at generating texts that evoke frames, they are not all equally good at this task. The results of analyzing

Table 2. How well the generated texts evoked the requested frames, by LLM and story-source. Note that correct texts might still contain hallucinated details (not be faithful to the source) despite evoking the right frame.

Model	Source	Correct	Faithful
GPT-4	Original	10/10	-
GPT-4	Bible	9/10	9/9
GPT-4	Sci-fi	8/10	5/8
Mistral-7B	Original	19/20	-
Mistral-7B	Bible	15/20	13/15
Mistral-7B	Sci-fi	13/20	6/13
Llama-2-7B	Original	18/20	-
Llama-2-7B	Bible	14/20	5/14
Llama-2-7B	Sci-fi	11/20	8/11
Yi-6B	Original	16/20	-
Yi-6B	Bible	7/20	6/7
Yi-6B	Sci-fi	16/20	10/16
Vicuna-7B	Original	16/20	-
Vicuna-7B	Bible	9/20	8/9
Vicuna-7B	Sci-fi	8/20	5/8

the quality of the texts are presented in Tables 1 and 2. Overall, we find that about 90% of the texts generated by GPT-4 evoke the correct frames, 78% for Mistral-7B, 72% for Llama-2-7B, 65% for Yi-6B, and only 55% for Vicuna-7B.

While the proprietary GPT-4 model is much larger in size, the other four models have comparable capacity and are all instruct-tuned; One major difference among the 7B models could be the amount of their training data, but this is not their only difference, and it would be interesting to further explore what causes the different performances in frame generation.

In many cases when the frame isn't evoked by the model's text, the reason has to do with *concept boundaries*. For example, while the annotators agreed that a 'dictator' and a 'strict father' have similarities, they didn't see them as the same. Interestingly, the original stories received better scores from the annotators than the quoted (Bible or sci-fi) texts.

A final observation is that the faithfulness of the generated texts (to the source) is a different axis than whether the frame is correctly evoked. Stated differently, about a *third* of the quoted Bible or sci-fi texts contained hallucinated details, while still evoking the right frame.

4.2 LLMs Can Zero-shot Recognize Frames (to Some Extent)

In the second set of experiments, we tested the capability of LLMs to recognize the frames implicit in the different texts in zero-shot settings. If one quizzes the LLMs about the definitions and characteristics of the frames in the study, they will give good answers. This is to be expected, since current LLMs have memorized information about many common topics [21]. But of course, being

Table 3. Zero-shot detection of SF and NP frames by Llama models. (The model was asked about the percentage that a story evokes these frames and the table counts responses with percentage ≥80%. The bold cells are as expected high; the underlined cells are unexpectedly high and suggest a misclassification.)

	Llama-3-70B		Llama-3-8B		Llama-2-7B	
Label	Z.SF	Z.NP	Z.SF	Z.NP	Z.SF	Z.NP
Strict Father	16 (**100%**)	5 (31%)	14 (**88%**)	13 (<u>81%</u>)	11 (69%)	13 (<u>81%</u>)
Nurturing Parent	3 (15%)	20 (**100%**)	2 (10%)	17 (**85%**)	13 (<u>65%</u>)	18 (**90%**)
Others/Control	17 (14%)	18 (15%)	10 (8%)	3 (3%)	38 (31%)	74 (<u>62%</u>)

able to describe a frame does not necessarily mean that an LLM can recognize it in practice, which is what we really want to assess.

In our initial setup, we prompted the LLMs openly to identify the top cognitive frames that were evoked by each text. However, a problem quickly emerged for this setup: a story may actually evoke multiple frames; For example, a father that checks their child's homework matches both the SF and NP frames. A story may also evoke a frame that isn't in our study, making evaluation difficult. Consequently, we designed this experiment as a *soft multi-label classification task* focused on the SF and NP frames. For reliability, we used only the texts that the annotators had deemed correct, and also excluded texts from the Vicuna model due to its lower quality (final n=16+20+120).[9]

We tested our zero-shot prompt on the Llama-2-7B-Chat model, as the 'average LLM' among our set. Due to the availability of the Llama-3-8B-Instruct and Llama-3-70B-Instruct [25] models at this time, we also decided to test the prompt on these improved models. The results are shown in Table 3.

The results show that the Llama-3-70B model can recognize the SF and NP frames in a zero-shot setting effectively. It recognizes all the true cases, and the overlap of the frames in the other cases is (as explained) reasonable. The differences between the model performances is, to our surprise, quite large. The zero-shot scores (evocation percentages) of Llama-3-70B and Llama-3-8B correlate at 0.72, while between Llama-3-70B and Llama-2-7B the correlation is a mere 0.19. In fact, according to Table 3, Llama-2-7B seems to be misclassifying a large number of the stories. This disparity raises intriguing questions about how frame recognition capabilities develop across model iterations and sizes which we leave for future work.

[9] An additional problem was that many of the NP and SF texts contained 'giveaway' words, e.g., the name of the frame was directly mentioned, or the parent was explicitly qualified as 'strict' or 'nurturing'. To avoid overt bias, we manually rephrased these texts and removed references to the words 'strict' and 'nurtur*'.

5 Frame Mechanics in the Hidden Representation

Having established that LLMs can both generate and recognize socio-political frames (especially in the Llama-3 models), and inspired by the research in mechanistic interpretability, we turn to the question of where frames are located inside the models.

Causal Tracing. As described in Section 2, we know from prior work that when we an LLM generates a text or completes a prompt about a frame, the information about the frame likely exists on the last subject token—which in our case means the last token encompassing the frame name. Furthermore, information about subject tokens are gradually enriched in the early feed-forward (MLP) layers, and from the middle of the network onwards passed via the attention heads to the very last token of the prompt.

We replicate the method known as 'causal tracing' to test the above, and basically localize where the model stores information about socio-political frames. Causal tracing isolates the "effect of individual states within the network while processing a factual statement" [24] while tracing the flow of information in the LLM. To start with, the model is given the prompt *"In the "XXX" frame, misbehavior is met with ..."* to complete. The prompt is correctly completed (based on Lakoff's [20] ideas) with `'punishment and discipline'` for the SF frame and `'empathy and understanding'` for the NP frame.

Causal tracing runs this "network multiple times, introducing corruptions to frustrate the computation, and finally restoring individual states in order to identify" the locations that restores the prediction [24]. In our case, we replace the 'subject tokens', that is the name of the frame, with Gaussian noise. This obviously leads to the generation of incomprehensible outputs. We then restore different parts of the hidden representation per token and layer (replacing them with the hidden representations saved after the initial correct completion), aiming to see when the correct output reappears.

The results are presented in Fig. 1. The x-axis represents the layers of the model (from 0 to 31), the y-axis lists the tokens in the input prompt, and the expected correct output token is shown on the last row (right). The color intensity represents the probability of the correct output token being restored when that specific hidden state is restored. In line with our hypothesis and the literature, we find that the information about the SF/NP frames can be restored at two points: the early layers (for the last subject token), and the later layers (for the last prompt token).

Sparse Probing. Based on the above results, we can now focus on probing the hidden representation specifically at layer 17.[10]

We use a binary logistic regression probe for our analysis due to its simplicity and ability to identifying salient dimensions without over-fitting with a complex

[10] Layer 17 is one of the first dark layers in the 'late site' of Fig. 1.

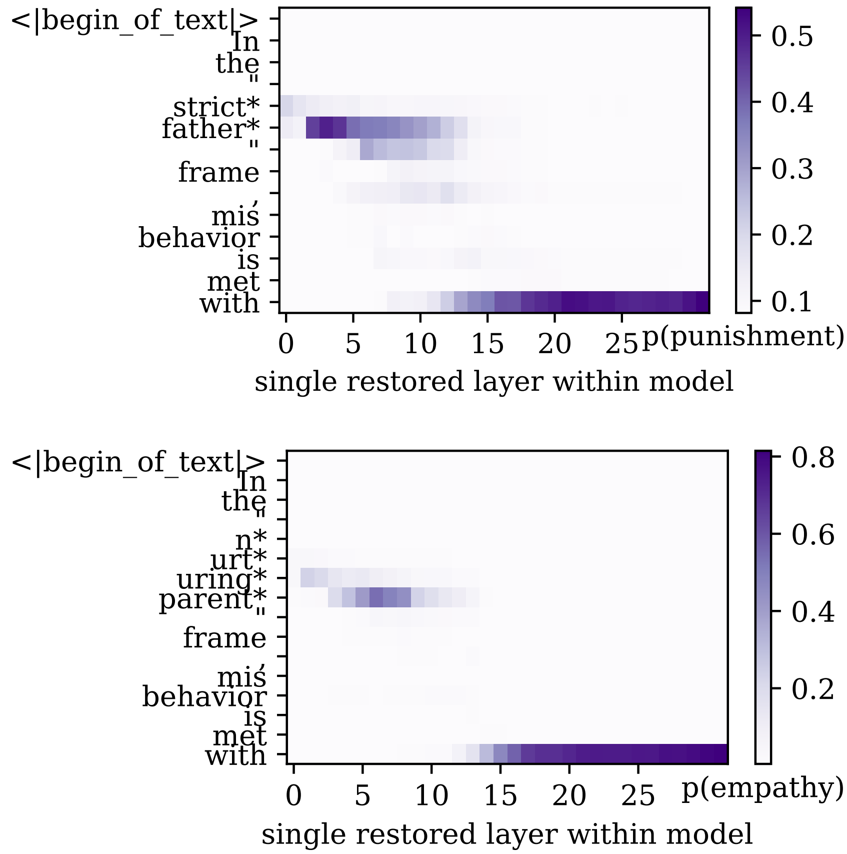

Fig. 1. Impact of restoring hidden state (for a single layer and token) on Llama-3-8B-Instruct's prediction ('punishment' or 'empathy') for the two prompts despite corrupting the subject tokens (SF/NP) in each prompt.

probe. To ensure our probe is really detecting the frames from the hidden representation, and not memorizing our dataset, we further limit the number of dimensions of the hidden representation that are used with recursive feature elimination to five and one (making it sparse).

We refine our dataset for the probing by preparing two binary classification tasks based on Table 3: SF vs. control and NP vs. control. Further, we only keep SF texts that evoke the SF frame according to both humans and machines, and NP texts that evoke the NP frame according to both humans and machines, and make sure the control group evokes neither frame. To have balanced classes, we expand the number of SF and NP texts by asking GPT4 to generate approx. 20 additional stories (final n≈39 per group; with 20% heldout).

Table 4. Statistics for a classifier probe that can identify texts that evoke the SF (or NP) frames from controls, using a limited number of hidden dimensions from the internal representation of layer 17 for the last token.

Frame	F1 score, 5 features	F1 score, 1 feature	Top Hidden Dim
Strict Father	0.93 (train) / 0.94 (test)	0.78 / 0.81	133
Nurturing Parent	0.90 / 0.88	0.80 / 0.88	529

The results for this experiment are shown in Table 4. What we find fascinating about these results is that we can tell the texts apart with an F1 score of around 80% with just 1 dimension out of the model's 4096 hidden dimensions.

The model may well be using shortcuts given the sample size–for instance, it might be detecting 'authority' or 'punishment' in the SF texts and 'care' in the NP texts. When we look at the top five (tokenized) words in the SF texts, they are[11]: father (all), discipline (22 times) , children (15 times), figure (10 times), and son (10 times). The top five words in the NP texts are: children (20 times), parent (19 times), child (17 times), love (15 times), and care (14 times). But every token will still have numerous dimensions in the internal representation space, and the fact that we can detect one distinctive dimension suggests that the SF/NP elements enforce each other.

6 Discussion & Conclusion

We found through our experiments and annotations that LLMs are generally fluent in a variety of socio-political frames. They can generate coherent original stories and or quote texts from other sources that evoke a frame—with all necessary sub-concepts of that frame.

It is perhaps not surprising, given the significance of frames in linguistics, psychology, and cognitive science, that machines that humans find understandable are fluent in frames. However, this still has important ethical and societal implications. Specifically, understanding how LLMs handle cognitive frames can have the positive consequence of demystifying the technology for a broader public. But it also suggests that the models may be able to create persuasive misinformation.

We further offered some initial investigations into the mechanics of the frame generation. Notably, we found that information about the frame persists throughout the generation process, specifically within the mid-layer hidden representations of the last prompt token.

This interdisciplinary study has been exploratory in nature, opening up areas for further research. We found, for one, that LLMs are not equal in their abilities with regard to cognitive frames, even at similar parameter sizes. This might be an area for further research, including testing more frames and model sizes, and

[11] This list excludes stop words and common words.

looking at the training data where possible to better understand how LLMs learn abstract socio-political frames.

Regarding the mechanics, it would be interesting to investigate if the presence of a frame can be decreased or increased—in line with similar efforts in AI safety research—to perhaps remove undesirable frames from a discourse.

In conclusion, our study reveals that LLMs possess a general fluency in generating and recognizing complex socio-political frames, opening up new avenues for both NLP researchers and social scientists. While this capability demonstrates the advanced nature of these models, it also underscores the need for careful consideration of their societal impact.

Acknowledgments. The authors are grateful to numerous colleagues who gave feedback on various stages of the research, including Jörg Pohle, Theresa Züger, and Freya Hewett. This study was funded by a grant from the German Federal Ministry of Science and Education (BMBF).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 01 AI, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., Dai, Z.: Yi: Open Foundation Models by 01.AI (Mar 2024). https://doi.org/10.48550/arXiv.2403.04652, http://arxiv.org/abs/2403.04652, arXiv:2403.04652 [cs]
2. Aghazadeh, E., Fayyaz, M., Yaghoobzadeh, Y.: Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2037–2050. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.144, https://aclanthology.org/2022.acl-long.144
3. Ali, M., Hassan, N.: A Survey of Computational Framing Analysis Approaches. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9335–9348. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.633, https://aclanthology.org/2022.emnlp-main.633
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623. FAccT '21, Association for Computing Machinery, New York, NY, USA (Mar 2021). https://doi.org/10.1145/3442188.3445922, https://doi.org/10.1145/3442188.3445922
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50

6. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (Jul 2020), http://arxiv.org/abs/2005.14165, rd: OpenAI/gpt3 paper
7. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (Mar 2023), https://lmsys.org/blog/2023-03-30-vicuna/
8. Chris Olah: Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases. Tech. rep. (Jun 2022), https://www.transformer-circuits.pub/2022/mech-interp-essay
9. Damstra, A., Boomgaarden, H.G., Broda, E., Lindgren, E., Strömbäck, J., Tsfati, Y., Vliegenthart, R.: What Does Fake Look Like? A Review of the Literature on Intentional Deception in the News and on Social Media. Journalism Studies **22**(14), 1947–1963 (Oct 2021). https://doi.org/10.1080/1461670X.2021.1979423, https://www.tandfonline.com/doi/full/10.1080/1461670X.2021.1979423, publisher: Routledge
10. Entman, R.M.: Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication **43**(4), 51–58 (Dec 1993). https://doi.org/10.1111/j.1460-2466.1993.tb01304.x, https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
11. Fillmore, C.J.: Chapter 10 Frame semantics. In: Cognitive Linguistics: Basic Readings, pp. 373–400. De Gruyter Mouton (Aug 2008). https://doi.org/10.1515/9783110199901.373, https://www.degruyter.com/document/doi/10.1515/9783110199901.373/html
12. FrameNet: What Is FrameNet, https://framenet.icsi.berkeley.edu/WhatIsFrameNet
13. Geva, M., Caciularu, A., Wang, K., Goldberg, Y.: Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 30–45. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.3, https://aclanthology.org/2022.emnlp-main.3
14. Geva, M., Schuster, R., Berant, J., Levy, O.: Transformer feed-forward layers are key-value memories. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 5484–5495. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.446, https://aclanthology.org/2021.emnlp-main.446
15. Gurnee, W., Tegmark, M.: Language models represent space and time. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=jE8xbmvFin
16. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B (Oct 2023). https://doi.org/10.48550/arXiv.2310.06825, http://arxiv.org/abs/2310.06825, arXiv:2310.06825 [cs]

17. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing, 3rd ed draft. Upper Saddle River, NJ: Prentice Hall (2024), https://web.stanford.edu/ jurafsky/slp3/
18. Kraut, R.: Plato. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2022 edn. (2022), https://plato.stanford.edu/archives/spr2022/entries/plato/
19. Lakoff, G.: Moral politics: what conservatives know that liberals do not. Mathematical Social Sciences **1**(33), 97–98 (1997)
20. Lakoff, G.: The ALL NEW Don't Think of an Elephant!: Know Your Values and Frame the Debate. Chelsea Green Publishing, White River Junction, Vermont, 10th anniversary edition edn. (Sep 2014)
21. Li, X.L., Kuncoro, A., Hoffmann, J., de Masson d'Autume, C., Blunsom, P., Nematzadeh, A.: A systematic investigation of commonsense knowledge in large language models. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11838–11855. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.812, https://aclanthology.org/2022.emnlp-main.812
22. Li, Y., Wang, S., Lin, C., Guerin, F., Barrault, L.: FrameBERT: Conceptual metaphor detection with frame embedding learning. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1558–1563. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.eacl-main.114, https://aclanthology.org/2023.eacl-main.114
23. Löbner, S.: Frames at the Interface of Language and Cognition. Annual Review of Linguistics **7**(1), 261–284 (2021). https://doi.org/10.1146/annurev-linguistics-042920-030620, https://doi.org/10.1146/annurev-linguistics-042920-030620, _eprint: https://doi.org/10.1146/annurev-linguistics-042920-030620
24. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems **36** (2022), arXiv:2202.05262
25. Meta: Llama 3 Model Card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
26. Mohler, M., Brunson, M., Rink, B., Tomlinson, M.: Introducing the LCC Metaphor Datasets. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4221–4227. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1668
27. Nikolaev, D., Baker, C., Petruck, M.R.L., Pad, S.: Adverbs, Surprisingly. In: Palmer, A., Camacho-collados, J. (eds.) Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). pp. 512–526. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.starsem-1.44, https://aclanthology.org/2023.starsem-1.44
28. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, Ilge, e.a.: GPT-4 Technical Report (Mar 2024). https://doi.org/10.48550/arXiv.2303.08774, http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs]

29. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2021)
30. Simon, F.M., Camargo, C.Q.: Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic'. New Media & Society **25**(8), 2219–2240 (Aug 2023). https://doi.org/10.1177/14614448211031908, https://doi.org/10.1177/14614448211031908, publisher: SAGE Publications
31. Touvron, H., Martin, L., Stone, K.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)
32. Vig, J.: Visualizing Attention in Transformer-Based Language Representation Models (Apr 2019), http://arxiv.org/abs/1904.02679, arXiv:1904.02679 [cs, stat]

A Additional Details on Prompts

A.1 Generation Prompts

The following prompts were used with `temperature=0.7` to generate stories:

- `Please write a short original story which evokes/invokes the "XXX" frame (max one paragraph).`
- `Please pick a short passage from the Bible which evokes/invokes the "XXX" frame (max one paragraph).`
- `Please describe a short story from a scifi novel or movie which evokes/invokes the "XXX" frame (max one paragraph).`

Importantly: no additional context or instructions were provided in the prompts about the frame or the task.

A.2 Zero-shot Prompts

The following prompts were used with `temperature=0` to recognize frames:

- `Can you tell me which major cognitive frames are evoked by the following text? (Please keep your answer strictly short and name max 5 frames with no explanation)`
- `What percentage does the following text evoke the "XXX" frame? (Please give just the percentage with no additional words)`

The following prompts were used to explore definitions and characteristics of the frames known to the (different) LLMs. The results helped pick the contrasting auto-completion sentences used in Section 5.

- `Please give a very short description of the "XXX" frame.`
- `Please describe key characteristics of the "XXX" frame.`