Do Robot Snakes Dream like Electric Sheep? Investigating the Effects of Architectural Inductive Biases on Hallucination

Anonymous ACL submission

Abstract

The growth in prominence of large language models (LLMs) in everyday life can be largely attributed to their generative abilities, yet some of this is also owed to the risks and costs associated with their use. On one front is their tendency to *hallucinate* false or misleading information, limiting their reliability. On another is the increasing focus on the computational limitations associated with traditional self-attention based LLMs, which has brought about new alternatives, in particular recurrent models, meant to overcome them. Yet it remains uncommon to consider these two concerns simultaneously. Do changes in architecture exacerbate/alleviate existing concerns about hallucinations? Do they affect how and where they occur? Through an extensive evaluation, we study how these architecture-based inductive biases affect the propensity to hallucinate. While hallucination remains a general phenomenon not limited to specific architectures, the situations in which they occur and the ease with which specific types of hallucinations can be induced can significantly differ based on the model architecture. These findings highlight the need for better understanding both these problems in conjunction with each other, as well as consider how to design more universal techniques for handling hallucinations.

1 Introduction

Large language models (LLMs) have rapidly emerged as a every-day tool in modern life (OpenAI, 2024), with many relying on their abilities to accomplish a variety of specific tasks. However, this opened up concerns relating to their propensity to *hallucinate* (Huang et al., 2023), with no concrete reasons for this behaviour (Dziri et al., 2022b; Rawte et al., 2023; Chen et al., 2024), hindering the ability to directly train LLMs that are consistently factual or able to explain themselves through their factual knowledge (Madsen et al., 2024a).

In parallel, as LLMs evolve and existing limitations are discovered, alternative architectures have become increasingly common and grow in popularity. In particular, Transformer LLMs (Vaswani et al., 2017) and linear sequence models (Gu et al., 2022; Gu and Dao, 2024) present contrasting methods of encoding sequences, with the Transformer using attention (Bahdanau et al., 2015) to form lossless representations of the context, while linear sequence models follow the recurrent neural network (Rumelhart et al., 1986; Jordan, 1986) in their use of a compressed state representation.

With the intensifying focus in both directions, a notable void exists in verifying how each can affect the others in conjunction. For example, existing works in hallucination detection and mitigation focus almost exclusively on Transformer-based models (Maynez et al., 2020; Longpre et al., 2021a; Guerreiro et al., 2023; Shi et al., 2023; Ji et al., 2023b; Farquhar et al., 2024; Wei et al., 2024), without extension to recurrent-style models, despite the use of a unified hidden representation potentially acting as an information bottleneck that can induce more common hallucinations. This lack of unified understanding on both topics prompts the need for a more explicit investigation.

In this work, we comprehensively explore the differences between pure attention LLMs and recurrent LLMs, specifically with respect to the propensity to hallucinate. Using a set of 20 different hallucination tasks, categorized into 6 groups that evaluate both faithfulness and factuality hallucinations, we evaluate across numerous open-source LLMs that range in scale from under 1B parameters to 70B parameters all the while covering a variety of different architecture choices such as self-attention, recurrent and hybrid models. We further evaluate across factors such as instruction-tuning, all to build a more comprehensive picture of architecture-specific phenomena with respect to hallucination. From this, we observe the following:

(1) Viewed very broadly over various different tasks and settings, neither Transformer-based nor recurrent/hybrid LLMs appear to induce hallucinations more often than others.

(2) However, shifting to individual tasks, it becomes evident that they result in disparate tendencies on specific tasks evaluating for unique criteria such as recalling long-tailed factual knowledge and falling into memorization traps, highlighting that model architecture may promote specific behavior that renders some types of hallucinations more common.

(3) Evaluating the effects of scale and instruction-tuning and how these changes affect performance, we observe that though factuality hallucinations remain consistently dependent on model size, recurrent/hybrid architectures are often more faithful at smaller sizes and observe significantly fewer faithfulness benefits from instruction-tuning and scaling compared to self-attention models.

These results highlight that while some hallucinations may be quite consistent across architectures, others are the direct results of specific model design choices that go into the LLM construction. This hints towards the need for more careful consideration on this front, as different techniques for addressing this problem can potentially be highly catered towards specific models, bringing to the forefront the need for better consideration of both these problems in the face of each other.

2 Related Work

Hallucinations in LLMs. Though various specific definitions exist, *hallucination* can broadly refer to "the generation of nonsensical information unfaithful to the provided source content" (Ji et al., 2023a). While this topic has become particularly important in order to ensure the safe and responsible use of language models, both classifying and quantifying the hallucinations is challenging. In particular, it is difficult to ascertain if the divergence occurs because of specific data heuristics (Lebret et al., 2016; Wiseman et al., 2017) or because of the innate lack of similarity between pre-training and downstream tasks (Rashkin et al., 2021), while measuring for hallucinations automatically introduces various biases (Reiter, 2018; Tian et al., 2020; Ganguli et al., 2022) that may not fully capture the scope of the errors. However, a variety of task-specific benchmarks (Li et al., 2020; Pagnoni et al., 2021; Zhou et al., 2021; Santhanam et al., 2022) have shown various LLMs to struggle with factual inconsistencies, highlighting a need to render them safer for every-day use.

LLM Architectures. General large language models architecture can be broadly thought to be composed of two components: token mixers which serves to model transformations between time steps (such as attention and recurrent layers) and channel mixers, such as multi-layer perceptrons and mixtures-of-experts (MoE) (Jacobs et al., 1991; Shazeer et al., 2017), which allow communication between different channels within a single time step. Accordingly, token mixing often forms a computational bottleneck in terms of time complexity while channel mixers consist of a memory bottleneck. Though the contemporary standard for token mixing remains self-attention, alternatives are becoming increasingly common as they begin to display more promise. These include the use of linear attention (Katharopoulos et al., 2020) to compensate for the quadratic memory complexity of vanilla self-attention, to new recurrent models (Gu and Dao, 2024; Team, 2024b) that function as a linear recurrent neural network but can process all elements of a sequence in parallel, and well as hybrid mixtures of recurrent mechanisms and attention (AI21, 2024; Dao and Gu, 2024) that have emerged as a meaningful competitor.

Architecture and Hallucination. Despite the growing research in new architectural components, whether or not how they affect hallucination has yet to be studied. While some works (Madsen et al., 2024b; Hu et al., 2024; Schimanski et al., 2024) have proposed modifications to either the learning/-generation pipeline as a way of reducing hallucinations, explicit proposals for hallucination reduction through structural modification have yet to be suggested. Additionally, though some works (Elhage et al., 2021; Fu et al., 2023; Lutati et al., 2023; Poli et al., 2024) have demonstrated self-attention to aptly solve synthetic tasks that form an essential component of language modelling, these fail to remain faithful on more realistic datasets, punctuating a major limitation of existing LLMs. Finally, though tangential work demonstrates that recurrent models may suffer from issues with information retention (Vardasbi et al., 2023), formally defining a link with hallucination remains necessary. These

issues lead to a lack of clarity on this front, accentuating the need for a formal investigation comparing hallucination alongside architectural paradigms. In particular, the choice of token mixer is important, as this is the component that governs how information is shared between different elements of the sequence to form a holistic representation. As such, an effective token mixer will adequately enable enough information from the context to propagate forward during the generation phase (Olsson et al., 2022; Arora et al., 2024), potentially enabling models to hallucinate less.

3 Background

Attention for Sequences. Vanilla self-attention as used in Transformers is powerful but costly. When provided an embedded text representation as a sequence of tokens $\boldsymbol{x} \in \mathbb{R}^{L \times d}$, each Transformer layer in the network applies a function

$$T_\ell(\boldsymbol{x}) = \text{FF}_\ell(A_\ell(\boldsymbol{x}) + \boldsymbol{x}) + A_\ell(\boldsymbol{x}) \qquad (1)$$

where A_ℓ is the self-attention of the ℓ-th layer and FF_ℓ is the following feed-forward network. Self-attention computes, for a token at position i in a sequence, a weighted average of the feature representations of all tokens (the values $\boldsymbol{V}_\ell$) in the sequence with a weight proportional to a similarity score between i (the query $\boldsymbol{Q}_\ell$ at position i) and the rest of the sequence (the keys $\boldsymbol{K}_\ell$). In particular, these can be computed for all positions in parallel

$$\begin{aligned} &\boldsymbol{Q}_\ell = \boldsymbol{x}\boldsymbol{W}_\ell^{\boldsymbol{Q}} \quad \boldsymbol{K}_\ell = \boldsymbol{x}\boldsymbol{W}_\ell^{\boldsymbol{K}} \quad \boldsymbol{V}_\ell = \boldsymbol{x}\boldsymbol{W}_\ell^{\boldsymbol{V}} \\ &A_\ell(\boldsymbol{x}) = \boldsymbol{V}_\ell' = \text{softmax}\big(\boldsymbol{Q}_\ell \boldsymbol{K}_\ell^T / \sqrt{d}\big)\boldsymbol{V}_\ell \end{aligned} \qquad (2)$$

providing the model a lossless representation of the complete past context. This can be seen as equivalent to search and retrieval within a database, where search is defined using query-key parameterizations and retrieval with value parameterization. In this setting, the database from which the information is being retrieved is equivalent to the model parameters, which store information from a training corpus that the weights $\boldsymbol{W}^{\boldsymbol{K}}, \boldsymbol{W}^{\boldsymbol{Q}}, \boldsymbol{W}^{\boldsymbol{V}}$ parameterizing the model attempt to mimic.

Multi-head attention, a variant of self-attention popularized by the Transformer (Vaswani et al., 2017), has become the dominant variant used in LLMs such as LLaMA (Touvron et al., 2023) and Gemma (Team, 2024a). Additional variants, such as multi-query attention from Falcon (Almazrouei et al., 2023) and grouped-query attention from Mistral (Jiang et al., 2023), adapt multi-head attention but remain build upon the same underlying principle of self-attention. Consequently, such methods are considered to fall under the family of self-attention token mixers.

From Transformers to Recurrent LLMs. One of the original concerns of Transformers was the quadratic complexity of attention, leading to large amounts of work focused on improving this bound as it could directly affect the ability to learn from long sequences. Instead of directly modifying attention (Katharopoulos et al., 2020; Kitaev et al., 2020; Choromanski et al., 2021), Gu et al. (2020) motivated a novel paradigm using state-space models (SSMs) from control theory. SSMs map an input $x(t) \in \mathbb{R}^d$ to an intermediate state $h(t) \in \mathbb{R}^n$ that is then projected to an output $y(t) \in \mathbb{R}^d$:

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \quad y(t) = \boldsymbol{C}h(t) + \boldsymbol{D}x(t)$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are all trainable parameters. Gu et al. (2021) use this paradigm to define a recurrent model to work on discrete signals, in which case the input can be regarded as discretized data sampled from a continuous signal with a step size Δ, for which the corresponding SSM is defined by:

$$\begin{aligned} &h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \quad y_t = \overline{\boldsymbol{C}}h_t + \overline{\boldsymbol{D}}x_t \\ &\overline{\boldsymbol{A}} = \frac{\big(I + \Delta\boldsymbol{A}/2\big)}{\big(I - \Delta\boldsymbol{A}/2\big)} \quad \overline{\boldsymbol{B}} = \frac{\Delta\boldsymbol{B}}{\big(I - \Delta\boldsymbol{A}/2\big)} \end{aligned}$$

and $\overline{\boldsymbol{C}} = \boldsymbol{C}$ ($\overline{\boldsymbol{D}}$ is equivalent to a residual connection and set to $\mathbf{0}$.) Thus

$$\overline{\boldsymbol{K}} = (\overline{\boldsymbol{C}}\overline{\boldsymbol{B}}, \overline{\boldsymbol{C}}\overline{\boldsymbol{A}}\overline{\boldsymbol{B}}, \ldots, \overline{\boldsymbol{C}}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}) \quad \boldsymbol{y} = \overline{\boldsymbol{K}} * \boldsymbol{x}$$

where $\overline{\boldsymbol{K}}$ is the SSM kernel. As $\boldsymbol{y}$ can be computed in $O(L \log L)$ with a Fast Fourier Transform (Cormen et al., 2009), the entire output can be computed in tandem based on the input, given the matrices that parameterize the system. Furthermore, setting $\boldsymbol{A}$ as a Hurwitz matrix, SSMs can preserve long-term dependency information and enable them to capture long-dependence information, overcoming a long-standing issue (Bengio et al., 1994) with prior recurrent models (Hochreiter and Schmidhuber, 1997; Cho et al., 2014).

Future works then modify this general structure for additional properties; Mamba (Gu and Dao, 2024) use input-dependent $\boldsymbol{B}$ and $\boldsymbol{C}$, enabling for

the model to filter information. LRU/Hawk (Orvieto et al., 2023; De et al., 2024) remove the discretization step from a continuous signal and instead learn discrete matrices for $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ directly. RWKV models (Peng et al., 2023, 2024) use a novel WKV token mixing operator that acts like an RNN. However, the token mixing paradigm for all such models shares one commonality, namely the use of hidden states from which token-level information must pass through in order to reach another token further in the future. Further interest in this topic has led to the creation of hybrid mixtures of recurrence and attention (Dao and Gu, 2024; De et al., 2024; AI21, 2024) to form new classes of models meant to capture the benefits of both.

Different time-mixing modules attempt to encode the history of tokens in a fixed representation. Recurrent architectures maintain this fixed representation and learn to update it during pre-training. Despite the variety of mechanisms and improvements to the representation according to the design of the module, they can fall short of attention (Vardasbi et al., 2023; Jelassi et al., 2024). One reason is that the amount of information that can be encoded (irrespective of length) is constrained by the size of hidden state (Arora et al., 2024). However, available architectures all converge during training on their respective pre-training and achieve comparable metrics. This then suggests that each architecture learns different aspects of the sequences that align with its own inductive biases. Self-attention architectures have the luxury to look at long range dependencies (albeit being costly). On the other hand, recurrent structures are constrained to make the predictions based on the recent context (which in practice can be long, but remains shorter than what self-attention can attend to). But the expressivity of the hidden representation as modeled by the time-mixers directly affects its performance on tasks (Sun et al., 2024) and the prior knowledge in the parameters of the architectures can be predisposed to favor the prediction of certain tokens. We model this effect on tasks under the lens of hallucination: faithfulness and factuality. Faithfulness expects the model to follow an instruction (that is largely recent), while factuality assumes the use of token-to-token relations learned during pre-training for a broader contextual representation. Following this, we ask two questions: (1) Do self-attention and recurrent architectures hallucinate differently? (2) Does instruction-tuning and/or size benefit the different classes of LLMs equally?

4 Experiments and Results

Models. For our comparison, we use various models (Table 2) that vary in size and architecture, with a particular focus on different time-mixing methods. These include self-attention models such as Pythia, LLaMA2/3 (Touvron et al., 2023), Falcon-7B (Almazrouei et al., 2023), Mistral-7B (Jiang et al., 2023), Gemma1/2 (Team, 2024a) and Mixtral-8x7B (Jiang et al., 2024). We also test recurrent models including Mamba (Gu and Dao, 2024; Dao and Gu, 2024), FalconMamba-7B and RWKV (Finch) (Peng et al., 2023), as well as hybrid models like RecurrentGemma (Team, 2024b) and Jamba (AI21, 2024). For models with instruction-tuned versions, we use both the base and instruction-tuned variant. Additional details are available in Appendix A.

Datasets. We evaluate on the Hallucination Leaderboard (Hong et al., 2024), consisting of tasks

1) **Closed-book Open-domain QA**♣: NQ-OPEN (Kwiatkowski et al., 2019), TRIVIAQA (Joshi et al., 2017), TRUTHFULQA (MC1, MC2, Generative) (Lin et al., 2022), POPQA (Joshi et al., 2017)
2) **Summarization**♠: XSUM (Narayan et al., 2018), CNN/DM (See et al., 2017)
3) **Reading Comprehension**♠: RACE (Lai et al., 2017), SQUADV2 (Rajpurkar et al., 2018), NQ-SWAP (Longpre et al., 2021b)
4) **Instruction Following**♠: MEMOTRAP (Liu and Liu, 2023), IFEVAL (Zhou et al., 2023)
5) **Hallucination Detection**: FAITHDIAL♠ (Dziri et al., 2022a), HALUEVAL♠ (QA, Summarization, Dialogue) (Li et al., 2023a), TRUEFALSE♣ (Azaria and Mitchell, 2023)
6) **Fact Checking**♣: FEVER (Thorne et al., 2018)

For tasks, a higher score (ranging from 0 to 100) indicates better performance. Tasks are further divided into two broader categories: *faithfulness* (♠), i.e. it evaluates whether an LLM generation adheres to the given source of information, and *factuality* (♣), i.e. it evaluates whether LLMs generate factually correct content according to world knowledge acquired during training.

4.1 Investigating Task Biases

At a broader scope that architecture does not appear to influence hallucination, both for task categories

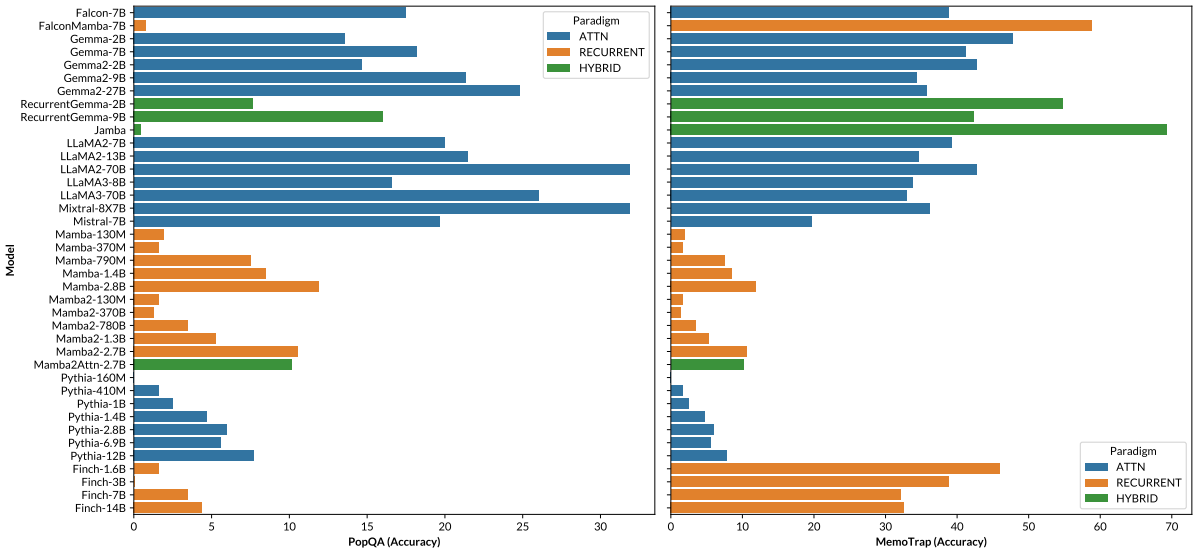


Figure 1: Performance on PopQA (left) and MemoTrap (right). Recurrent and hybrid models significantly outperform similar pure attention-based alternatives on MemoTrap, but the opposite is true on PopQA.

and factuality/faithfulness. For example, with like-sized models falling within different categories of their time mixers (self-attention, recurrent layers, hybrid), mean performance on tasks within each grouping is not significantly different nor do they show a particular ordering[1]. However, with specific tasks, differences start to present themselves, which we describe in our following observations.

> **O1**: Recurrent models struggle with rare knowledge.

Interestingly, recurrent and hybrid models all significantly underwhelm on PopQA, a task that tests for uncommon (thin-tailed) factual knowledge (left plot in Figure 1). For example, FalconMamba achieves an exact-match (EM) score of 0.7 compared to 17.5 by a similar Falcon. Similarly, 2B and 9B RecurrentGemma achieve 7.6 and 16.0 EM respectively, compared to 14.6/21.3 EM for 2B/7B Gemma and 13.6/18.2 by 2B/9B Gemma2, while Jamba (0.4 EM) performs worse than a similarly sized Mixtral-8x7B (31.9 EM).

Previous work (Vardasbi et al., 2023) suggests that hidden states can become dense and difficult to extract information from in the future. Although this is relevant for hallucinations, these concerns are unlikely to apply in this setting given that the prompt consists of only a question and the amount of information to store within the hidden representation is unlikely to lead to it becoming highly dense. Additional works (Dao and Gu, 2024) have instead posited that using recurrent models with additional mechanisms enabling input filtering can enhance performance on information-dense data such as language; however, this is again unlikely to be the case due to the general simplicity of the task. Yet this result is significant as it indicates recurrent models, despite the task's simplicity, struggle when compared to attention-based equivalents, suggesting a direct link to the token mixing that is utilized. The reason may instead be that recurrent/hybrid models do not learn by explicitly memorizing and acting as a retrieval system (as discussed in Section 3). In this sense, despite potentially observing low-tail information, the learning mechanism may not have led to said information being stored within parameters, making recurrent/hybrid models more likely to produce incorrect information in such settings with higher degrees of uncertainty.

> **O2**: Recurrent models can better follow contexts by relying less on memory.

There are also cases where recurrent models consistently outperform attention equivalents, namely on MemoTrap (right plot in Figure 1). Here, the LLM is prompted to complete a well-known proverb with an ending that deviates from the commonly used ending, testing for over-reliance on knowledge memorized from a training corpus. This suggests that although recurrent models might

[1] Explicit numbers in Table 3, 4 and 5 in Appendix B.

Model Name	Hallu. Detection	Instr. Following	Closed-Book QA	Reading Comp.	Sum.	Fact-Checking	FAITHFULNESS	FACTUALITY
Attention-Only Models								
Gemma-2B	49.65 (↑ 11.51)	30.25 (↑ 13.84)	38.20 (↓ -6.92)	38.96 (↓ -13.96)	24.93 (↓ -4.53)	54.14 (↑ 1.60)	35.85 (↑ 1.92)	32.56 (↓ -4.56)
Gemma-7B	53.81 (↓ -5.67)	31.94 (↑ 6.79)	27.43 (↓ -4.13)	31.35 (↓ -5.83)	20.63 (↑ 1.04)	43.28 (↓ -5.22)	36.48 (↑ 2.71)	44.50 (↓ -6.04)
Gemma2-2B	58.33 (↑ 0.64)	26.73 (↑ 10.90)	29.59 (↑ 4.83)	32.95 (↓ -5.26)	16.04 (↓ -0.47)	39.57 (↑ 18.96)	35.61 (↑ 0.65)	34.96 (↓ -0.98)
Gemma2-9B	65.64 (↑ 6.31)	23.76 (↑ 30.07)	39.36 (↑ 3.13)	39.74 (↓ -0.70)	21.38 (↓ -3.35)	62.06 (↑ 6.47)	40.56 (↑ 7.55)	46.85 (↑ 3.13)
Gemma2-27B	62.03 (↑ 12.43)	26.62 (↑ 33.58)	47.19 (↑ 2.86)	43.04 (↑ 5.02)	28.92 (↓ -0.74)	68.28 (↑ 1.00)	41.54 (↑ 13.02)	53.92 (↑ 2.36)
LLaMA2-7B	53.63 (↑ 4.30)	28.94 (↑ 8.99)	37.64 (↓ -1.89)	27.58 (↑ 3.74)	25.19 (↑ 0.69)	51.38 (↑ 5.88)	35.14 (↑ 4.01)	42.51 (↓ -0.12)
LLaMA2-13B	67.80 (↓ -3.68)	26.57 (↑ 9.72)	39.65 (↓ -0.65)	32.17 (↓ -3.64)	27.33 (↓ -0.23)	62.35 (↓ -1.64)	40.83 (↑ 0.88)	46.27 (↓ -0.49)
LLaMA2-70B	62.34 (↑ 10.78)	30.24 (↑ 17.11)	47.85 (↓ -2.53)	39.48 (↓ -8.21)	28.00 (↓ -0.55)	66.63 (↓ -1.45)	41.74 (↑ 5.82)	53.69 (↓ -2.34)
LLaMA3-8B	60.76 (↑ 10.12)	22.06 (↑ 26.76)	41.55 (↑ 1.30)	33.52 (↑ 3.62)	26.62 (↓ -1.60)	60.84 (↑ 4.82)	37.60 (↑ 10.09)	48.14 (↑ 1.60)
LLaMA3-70B	71.78 (↑ 8.69)	21.59 (↑ 25.60)	48.07 (↑ 3.43)	46.38 (↓ -7.73)	28.91 (↓ -1.68)	69.57 (↑ 0.99)	45.92 (↑ 6.23)	54.80 (↑ 2.81)
Mistral-7B	60.48 (↑ 3.36)	28.39 (↑ 16.85)	41.24 (↑ 5.21)	32.03 (↑ 1.92)	26.77 (↓ -0.77)	58.59 (↑ 6.78)	38.47 (↑ 4.67)	47.42 (↑ 4.17)
Mixtral-8x7B	73.51 (↓ -0.15)	26.84 (↑ 17.60)	48.99 (↑ 4.33)	40.66 (↓ -3.90)	27.91 (↓ -1.07)	68.35 (↑ 0.49)	46.16 (↑ 3.89)	55.15 (↑ 3.50)
Falcon-7B	52.40 (↓ -2.90)	25.54 (↑ 12.62)	33.03 (↓ -4.42)	29.18 (↓ -3.64)	20.70 (↓ -1.00)	46.91 (↓ -7.71)	34.91 (↓ -0.71)	36.92 (↓ -3.75)
Recurrent and Hybrid Models								
RecurrentGemma-2B	52.88 (↑ 2.33)	33.43 (↑ 3.12)	27.36 (↓ -1.85)	30.66 (↓ -1.77)	18.41 (↑ 2.86)	42.97 (↑ 3.63)	36.55 (↓ -0.14)	31.46 (↑ 1.39)
RecurrentGemma-9B	55.75 (↓ -1.67)	31.67 (↑ 12.96)	36.79 (↑ 2.03)	36.57 (↓ -6.66)	22.99 (↑ 1.84)	51.25 (↑ 13.36)	37.61 (↓ -0.17)	42.96 (↑ 3.28)
Jamba	57.66 (↓ -2.90)	43.36 (↓ -5.76)	39.50 (↓ -0.98)	33.00 (↓ -5.88)	23.72 (↓ -15.48)	59.88 (↓ -2.60)	39.80 (↓ -6.91)	45.95 (↓ -0.91)
FalconMamba-7B	55.80 (↑ 0.73)	42.97 (↓ -1.19)	39.94 (↓ -0.43)	23.76 (↓ -0.85)	23.89 (↓ -0.09)	61.85 (↓ -1.80)	35.92 (↓ -0.26)	47.02 (↓ -0.38)

Table 1: Changes in performance from the use of instruction-tuning. While factuality does not exhibit a specific trend, faithfulness is shown to improve within attention-based models but not recurrent or hybrid models.

struggle with long-tail knowledge, they are less prone to ignoring contextual cues not stored within parametric memory, or they more model the dynamics based on the context as opposed to acting as a retrieval system. Noticeable drops exist from FalconMamba (58.8) to Falcon (38.9) and from RecurrentGemma 2B/9B to Gemma2 2B/9B (53.7/43.4 vs 42.7/34.4), suggesting that attention can lead to more hallucinations. It further reveals a benefit of recurrent layers; by not memorizing information directly within parameters, there may be a reduced tendency to repeat previous observations (Jelassi et al., 2024) and greater focus on the context.

O3: Scale is required for emergent qualities.

However, some scale to the data and model remain important. Note that Mamba, Finch and Pythia fail to draw the same distinctions as the other models despite sharing the same vocabulary and training data. However, given the scale of training for these (300B tokens compared to >1T tokens for other models) and the fact that the discussed tasks relate directly to information memorization, it is possible that these phenomena fail to emerge at this data scale. This is particularly evidenced by how these models perform better with size on MEMOTRAP, which is designed in a way such that larger models should normally perform worse.

4.2 On the role of Instruction-Tuning.

The use of instruction-tuning appears to show inconsistent improvements across categories (Table 1) and tasks[2]. There exists no strict pattern for overall hallucination; on some tasks, its use is effective for all (ex. TRUTHFULQA) or no (ex. TRIVIAQA) models. Smaller trends also exist, such as instruction-following observing the strongest gains.

However, changes in faithfulness are generally positive for pure attention models (shown in teal) and negative for models with recurrent layers (shown in magenta), whereas changes in factuality show no consistent trends. This suggests that instruction-tuning for LLMs, as initially proposed with self-attention based models, enables particular properties which are no longer relevant after the addition of recurrent layers. Their inductive biases therefore can have a particular effect on specific forms of hallucinations emerging, with instruction-tuning being a manner to mitigate these.

Recalling the links between attention and retrieval from Section 3, the process of instruction-tuning can be understood as shifting the projection weights towards a new space through the additional fine-tuning (which is of itself simply an additional step of language modeling). This leads to the overwriting of some of the pre-training corpus with that used for fine-tuning, effectively replacing the keys and values from the underlying retrieval storage. While effects are ambiguous on factuality, as some potentially relevant facts can be removed, faithfulness often improves from this process, as the fine-tuning corpus is more likely to contain information relevant to context following and thereby make it easier for the attention mechanism to retrieve rel-

[2]See Figure 5.

evant information for such a purpose. In addition to highlighting how enhancing the ability to follow instructions does not necessarily lead to factually correct results, another conclusion stemming from the fact that recurrent models generally observe no benefits in faithfulness after instruction-tuning is that such models' inductive biases render them inherently more faithful or the process of making them more faithful is distinct from self-attention.

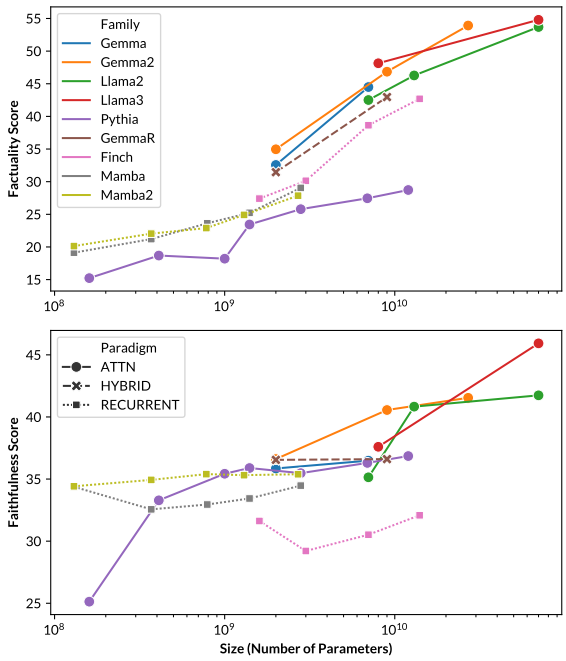


Figure 2: Changes in factuality (top) and faithfulness (bottom) as models are increased in size. Score range between 0 and 100. Factuality always increases with the number of parameters, however faithfulness increases are only meaningful for pure-attention models.

4.3 Impact of Model Size on Hallucinations.

With the large amount of prior work (Kaplan et al., 2020; Wei et al., 2022; Hoffmann et al., 2022) concluding that model size can play a significant role in model reasoning abilities, we provide this additional axis of variation in Figure 2.

Immediately, we observe that increased size leads to improvements in factuality regardless of model. This is partially expected, considering the range of work that has suggested links between size and the amount of factual knowledge that can be encoded within parametric space. However, faithfulness is different. Recurrent and hybrid architectures tend to saturate in terms of faithfulness at a small size whereas pure attention models do not. For example, Mamba models stagnate in faithfulness at around 130M, whereas Pythia continues to increase in this score up to billions of parameters. Finch and RecurrentGemma also show this while attention-based models do not. Another observation is the fact that even the smallest Mamba are nearly as faithful as LLaMA2-7B and Pythia-1.4B and despite models eventually catching up in terms of faithfulness, having such small models perform on par indicates a meaningful manner in which their hallucinations may differ. To further illustrate this, we provide an additional depiction of a performance to scale metric, defined as

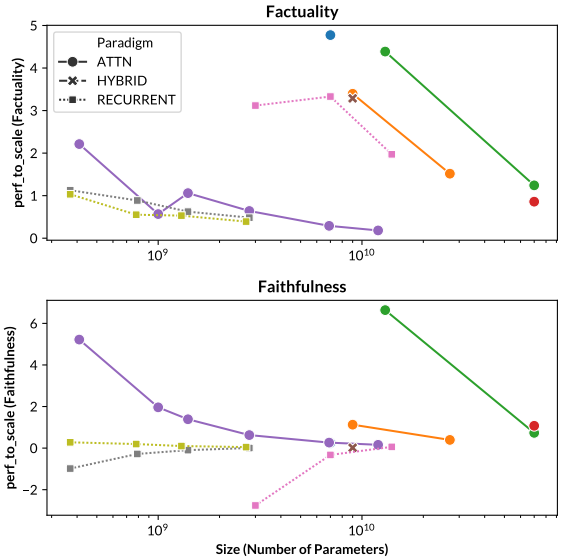


Figure 3: Performance to scale values for factuality and faithfulness. Colors differentiate model families while shapes differentiate the model type. Factuality improves with model size (indicated by values >0) and the model type does not have a link with this relative improvement. Recurrent/hybrid models show low values for faithfulness, indicating that size increases generally do not benefit them, unlike attention models.

$$\texttt{perf_to_scale}(m, \mathcal{M}, \mathcal{B}) = \frac{m(\mathcal{M}) - m(\mathcal{B})}{P(\mathcal{M})/P(\mathcal{B}) - 1}$$

where $m(\cdot)$ is a metric of interest, $\mathcal{M}$ is a model and $\mathcal{B}$ is a base model of the family as $\mathcal{M}$. $P(\cdot)$ counts the number of parameters in a given model. Namely, this measures the relative increase in m by linear scaling the model, with a lower value indicating that linearly increasing the number of parameters yields no changes while larger values indicate scaling leads to significant improvements. More positive values indicate that the score increases greatly with linear increases in size and negative values the opposite. Consistently negative or near-zero values hence mean that size has little effect

on the score. Figure 3 depicts the performance to scale changes for factuality and faithfulness, where it becomes apparent that increasing the number of parameters in hybrid and recurrent models yields a significantly lower increase in faithfulness compared to pure attention models.

This suggests that recurrent layers enable an inherent ability to follow contexts not improved by model size. Their inductive biases can be perceived as encouraging context following, supported by how the hidden state incorporates information from the context with the current input to control the generation. This perspective aligns these results with our instruction-tuning observations, where such models do not exhibit a clear benefit compared to their attention-based counterparts. In particular, we can interpret this as indicating that the base model learns to perform this output control even without direct instruction-tuning; hence applying these techniques is ineffective as we previously observed.

5 Discussion

Hallucination Mitigation for Different Architectures. Hallucinations motivate the need to both detect and mitigate them, particularly when they can escalate to have adverse effects. Existing techniques may be insufficient, as they contain potential implicit biases that can fundamentally conflict with the manners in which hallucinations may differ between models. Facts such as sequence models saturating in faithfulness in a manner not resolved through instruction-tuning indicates that existing mitigation techniques can differ in their effectiveness due to a bias towards specific architectures.

This means that current techniques also further deserve a look, in particular in terms of their effectiveness across the different inductive biases of models. While some methods evidently might be specifically catered to specific models, e.g. using attention (Li et al., 2023b; Zhang et al., 2024; Chuang et al., 2024), other methods may implicitly also possess biases due to the underlying assumptions being made or specific requirements that may be unevenly addressed through different architectures, such as the reliance on data-refinement or additional context (Shi et al., 2024).

Inherent Faithfulness and Factuality. Designing inherently faithful (Herman, 2019; Wiegreffe and Pinter, 2019) or factual models remains an important issue. While much work suggests attention-based LLMs lack an inherent tendency to be faithful (Jacovi and Goldberg, 2020; Wiegreffe and Marasovic, 2021), methods have been designed to render models more interpretable (Madsen et al., 2024b) for specifically evaluating this. Similarly, some of these methods have been shown to also adapt to Mamba models (Sharma et al., 2024). However, while interpretability is useful for evaluating models as being faithful or factual, it does not ensure an inherent tendency towards being either.

Nevertheless, these results hint that architectures can provide an avenue through which desiderata can be attained. SSMs present an option which in particular appears suited for settings which require utilizing information provided within a context, given their use within control theory to model system dynamics. With language, this can be analogous to using the fixed-size representation of the previous context for control, which may mean that such models can remain more faithful as the context remains the source that guides the generation. In comparison, attention-based models have been shown to act as retrieval systems that retrieve tokens based on high similarity with memorized information; hence, when a large difference exists between the context and what has been stored from training, breakdowns in faithfulness can occur.

6 Conclusion

Are hallucinations the direct result of how token-mixing takes place across time? We study whether different inductive biases of LLM architectures can increase this propensity by observing how hallucinations change across numerous tasks. Patterns emerge demonstrating that some architectural biases can lead to either improvements or degradation in performance compared to others, both for individual tasks and categorizations. Additionally, some model types lead to inherent behaviour when comparing the effects of instruction-tuning and scaling model size. In sum, model-specific inductive biases can have a direct effect on the type of hallucinations they are faced with. However, the data used and the specific model construction can also affect learning. We hope that future work can build on our findings by exploring both the types of hallucinations that can occur and techniques meant to mitigate them and how to design more universal techniques to make models more robust or reliable.

7 Limitations

Evaluating faithfulness and factuality. A major limitation existing in hallucination detection and mitigation research is the lack of metrics that provide explicit information regarding hallucination. While the development of such metrics is an active area of research, limitations still exist, such as biases when involving additional models within the evaluation process.

Limits on tasks. Another limitation of this work is the non-exhaustive set of tasks and domains in which we measure hallucinations. Due to the exhaustive ways in which hallucinations can be measured, we limit ourselves to tasks that are well motivated and frequently used in practice.

8 Ethical Concerns

This paper provides an analysis on the effects of different architectural inductive biases on the propensity to hallucinate within large language models. As such, mistakes in methodology can lead to unsupported confidence or skepticism regarding their performance with the explored task or related ones. While skepticism may not be an ethical issue, unsupported confidence can be problematic. However, the overall message is that all LLMs have specific tasks and settings in which they will display a greater propensity to hallucinate, hence we do not believe that the ideas expressed in this work will explicitly lead to unsupported confidence.

References

AI21. 2024. Introducing jamba: Ai21's groundbreaking ssm-transformer model.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Re. 2024. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2024. Is bigger and deeper always better? probing llama across scales and layers. *Preprint*, arXiv:2312.04333.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *Preprint*, arXiv:2407.07071.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms, Third Edition*, 3rd edition. The MIT Press.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. 2024. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *Preprint*, arXiv:2402.19427.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Daniel Y Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W Thomas, Benjamin Frederick Spector, Michael Poli, Atri Rudra, and Christopher Re. 2023. Monarch mixer: A simple sub-quadratic GEMM-based architecture. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. 2020. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, pages 1474–1487. Curran Associates, Inc.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*.

Albert Gu, Isys Johnson, Karan Goel, Khaled Kamal Saab, Tri Dao, Atri Rudra, and Christopher Re. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *Advances in Neural Information Processing Systems*.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Bernease Herman. 2019. The promise and peril of human evaluation for model interpretability. *Preprint*, arXiv:1711.07414.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard – an open effort to measure hallucinations in large language models. *Preprint*, arXiv:2404.05904.

Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma, and Irwin King. 2024. Mitigating large language model hallucination with faithful finetuning. *Preprint*, arXiv:2406.11267.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and eran malach. 2024. Repeat after me: Transformers are better than state space models at copying. In *Forty-first International Conference on Machine Learning*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023b. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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. *Preprint*, arXiv:2401.04088.

Michael I. Jordan. 1986. Serial order: a parallel distributed processing approach. Technical report, University of California, San Diego: Institute for Cognitive Science.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu. 2020. Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 97–106, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Alisa Liu and Jiacheng Liu. 2023. The memotrap dataset.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021a. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021b. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shahar Lutati, Itamar Zimerman, and Lior Wolf. 2023. Focus your attention (with adaptive IIR filters). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12538–12549, Singapore. Association for Computational Linguistics.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024a. Are self-explanations from large language models faithful? *Preprint*, arXiv:2401.07927.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2024b. Faithfulness measurable masked language models. In *Forty-first International Conference on Machine Learning*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. 2023. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr. au2, Jiaju Lin, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Cahya Wirawan, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *Preprint*, arXiv:2404.05892.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. 2024. Mechanistic design and scaling of hybrid architectures. *Preprint*, arXiv:2403.17844.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,

pages 704–718, Online. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2022. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *Preprint*, arXiv:2110.05456.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in mamba. In *First Conference on Language Modeling*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. 2024. Learning to (learn at test time): Rnns with expressive hidden states. *Preprint*, arXiv:2407.04620.

Gemma Team. 2024a. Gemma.

Gemma Team. 2024b. Recurrentgemma: Moving past transformers for efficient open language models. *Preprint*, arXiv:2404.07839.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2020. Sticking to the facts: Confident decoding for faithful data-to-text generation. *Preprint*, arXiv:1910.08684.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ali Vardasbi, Telmo Pessoa Pires, Robin Schmidt, and Stephan Peitz. 2023. State spaces aren't enough: Machine translation needs attention. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 205–216,

Tampere, Finland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers. *Preprint*, arXiv:2402.10412.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Technical Implementation Details

A.1 Models and Datasets

All models and datasets used in this paper are public and directly available on the HuggingFace Hub.

A.2 Experimental Setup

Experiments were run using the Language Model Evaluation Harness (Gao et al., 2024).

A.3 Datasets

We evaluate on the Hallucination Leaderboard (Hong et al., 2024), consisting of tasks

1) **Closed-book Open-domain QA**♣: NQ-OPEN (Kwiatkowski et al., 2019), TRIVIAQA (Joshi et al., 2017), TRUTHFULQA (MC1, MC2, Generative) (Lin et al., 2022), POPQA (Joshi et al., 2017)
2) **Summarization**♠: XSUM (Narayan et al., 2018), CNN/DM (See et al., 2017)
3) **Reading Comprehension**♠: RACE (Lai et al., 2017), SQUADV2 (Rajpurkar et al., 2018), NQ-SWAP (Longpre et al., 2021b)
4) **Instruction Following**♠: MEMOTRAP (Liu and Liu, 2023), IFEVAL (Zhou et al., 2023)
5) **Hallucination Detection**: FAITHDIAL♠ (Dziri et al., 2022a), HALUEVAL♠ (QA, Summarization, Dialogue) (Li et al., 2023a), TRUEFALSE♣ (Azaria and Mitchell, 2023)
6) **Fact Checking**♣: FEVER (Thorne et al., 2018)

For tasks, a higher score (ranging from 0 to 100) indicates better performance. Tasks are further divided into two categories: *faithfulness* (♠) hallucinations, i.e. whether an LLM generation adheres to the given source of information, and *factuality* (♣) hallucinations, i.e. whether LLMs generate factually correct content according to world knowledge based on knowledge acquired during training.

A.4 Models and Baselines

For our comparison, we use various models (Table 2) that vary in size and architecture, with a particular focus on different time-mixing methods. These include:

- All Pythia models
- All publicly available Mamba (Gu and Dao, 2024; Dao and Gu, 2024) models

- All LLaMA2 and LLaMA3 (Touvron et al., 2023) models
- Falcon-7B (Almazrouei et al., 2023) and FalconMamba-7B
- Mistral-7B (Jiang et al., 2023)
- All Gemma (Team, 2024a) models
- RWKV (Peng et al., 2023) models, namely Finch (RWKV-v6)
- RecurrentGemma (Team, 2024b)
- Mixtral-8x7B (Jiang et al., 2024)
- Jamba (AI21, 2024)

For models with instruction fine-tuned versions, we use both the base version as well as the instruction fine-tuned variant.

A.5 Computing Resources Used

All results were obtained using a server of 8 NVIDIA V100 32GB or 4 NVIDIA RTX A6000 48GB GPUs. The `accelerate` package was used for model sharding in instances where a single GPU was insufficient to store the entire model.

B Complete Results

B.1 Instruction Tuning Changes

Model	Public Link	HuggingFace Model
Pythia-160M	EleutherAI/pythia-160m	✔
Pythia-410M	EleutherAI/pythia-410m	✔
Pythia-1B	EleutherAI/pythia-1b	✔
Pythia-1.4B	EleutherAI/pythia-1.4b	✔
Pythia-2.8B	EleutherAI/pythia-2.8b	✔
Pythia-6.9B	EleutherAI/pythia-6.9b	✔
Pythia-12B	EleutherAI/pythia-12b	✔
Mamba-130M	state-spaces/mamba-130m	✘
Mamba-370M	state-spaces/mamba-370m	✘
Mamba-790M	state-spaces/mamba-790m	✘
Mamba-1.4B	state-spaces/mamba-1.4b	✘
Mamba-2.8B	state-spaces/mamba-2.8b	✘
Mamba2-130M	state-spaces/mamba2-130m	✘
Mamba2-370M	state-spaces/mamba2-370m	✘
Mamba2-780M	state-spaces/mamba2-780m	✘
Mamba2-1.3B	state-spaces/mamba2-1.3b	✘
Mamba2-2.7B	state-spaces/mamba2-2.7b	✘
Mamba2Attention-2.7B	state-spaces/mamba2attn-2.7b	✘
RecurrentGemma-2B	google/recurrentgemma-2b	✔
RecurrentGemma-2B (IT)	google/recurrentgemma-2b-it	✔
RecurrentGemma-9B	google/recurrentgemma-9b	✔
RecurrentGemma-9B (IT)	google/recurrentgemma-9b-it	✔
Gemma-2B	google/gemma-2b	✔
Gemma-2B (IT)	google/gemma-2b-it	✔
Gemma-9B	google/gemma-9b	✔
Gemma-9B (IT)	google/gemma-9b-it	✔
Gemma2-2B	google/gemma2-2b	✔
Gemma2-2B (IT)	google/gemma2-2b-it	✔
Gemma2-9B	google/gemma2-9b	✔
Gemma2-9B (IT)	google/gemma2-9b-it	✔
Gemma2-27B	google/gemma2-27b	✔
Gemma2-27B (IT)	google/gemma2-27b-it	✔
Falcon-7B	tiiuae/falcon-7b	✔
Falcon-7B (IT)	tiiuae/falcon-7b-instruct	✔
FalconMamba-7B	tiiuae/falcon-mamba-7b	✔
FalconMamba-7B (IT)	tiiuae/falcon-mamba-7b-instruct	✔
Mistral-7B	mistralai/Mistral-7B-v0.3	✔
Mistral-7B (IT)	mistralai/Mistral-7B-Instruct-v0.3	✔
Mixtral-8x7B	mistralai/Mixtral-8x7B-v0.1	✔
Mixtral-8x7B (IT)	mistralai/Mixtral-8x7B-Instruct-v0.1	✔
Jamba	ai21labs/Jamba-v0.1	✔
Jamba (IT)	ai21labs/AI21-Jamba-1.5-Mini	✔
LLaMA2-7B	meta-llama/Llama-2-7b-hf	✔
LLaMA2-7B (IT)	meta-llama/Llama-2-7b-hf	✔
LLaMA2-13B	meta-llama/Llama-2-13b-hf	✔
LLaMA2-13B (IT)	meta-llama/Llama-2-13b-hf	✔
LLaMA2-70B	meta-llama/Llama-2-70b-hf	✔
LLaMA2-70B (IT)	meta-llama/Llama-2-70b-hf	✔
LLaMA3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B	✔
LLaMA3-8B (IT)	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct	✔
LLaMA3-70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B	✔
LLaMA3-70B (IT)	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct	✔

Table 2: Models used and public links to their weights.

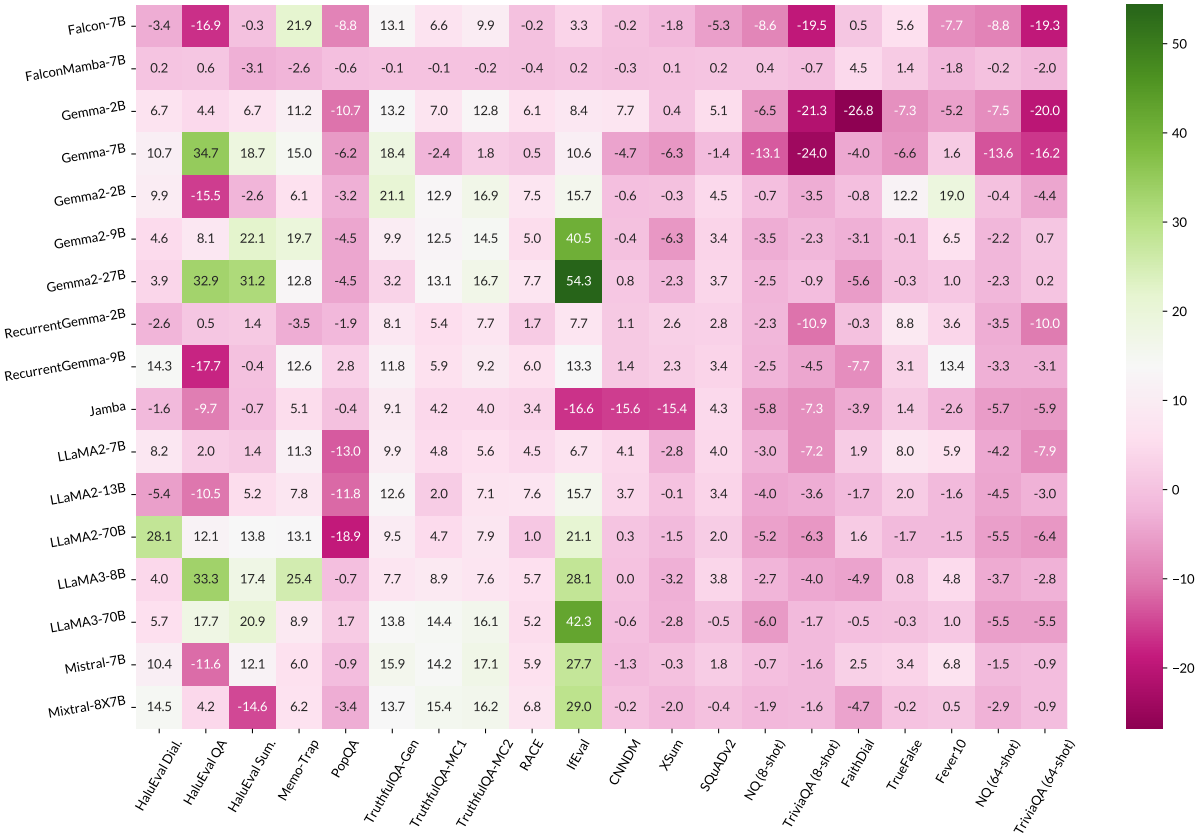
Figure 4: Performance of various base models on tasks within the Hallucination Leaderboard.

Name	HaluEval Dial.	HaluEval QA	HaluEval Sum.	Memo-Trap	PopQA	TruthfulQA-Gen	TruthfulQA-MC1	TruthfulQA-MC2	RACE	IfEval	CNNDM	XSum	SQuADv2	NQ (8-shot)	TriviaQA (8-shot)	FaithDial	TrueFalse	Fever10	NQ (64-shot)	TriviaQA (64-shot)
Falcon-7B	41.8	46.6	49.0	38.9	17.5	25.0	22.4	34.3	37.4	12.2	20.6	20.8	26.0	22.9	58.7	66.6	58.1	46.9	24.3	59.3
FalconMamba-7B	49.9	39.5	42.5	58.8	0.7	49.2	37.8	53.4	43.0	27.2	25.1	22.7	24.4	26.0	64.1	58.2	88.8	61.8	27.8	60.4
Finch-1.6B	47.7	51.0	50.3	45.9	1.6	26.1	22.2	36.7	35.6	16.3	6.6	16.9	11.0	10.3	36.7	63.2	53.1	39.5	10.9	37.0
Finch-3B	44.6	49.7	53.3	38.8	0.0	27.1	23.6	37.0	35.3	16.8	0.0	0.1	18.1	13.9	45.6	64.3	54.4	39.3	14.6	45.9
Finch-7B	52.2	49.4	48.6	32.1	3.4	38.7	27.4	42.2	35.5	19.0	0.7	10.2	27.3	21.5	60.0	60.4	64.6	45.3	23.2	60.2
Finch-14B	61.5	52.7	50.2	32.6	4.4	40.8	31.6	45.4	34.8	20.7	0.3	3.4	31.6	24.9	63.5	64.2	74.6	52.0	25.7	64.1
Gemma-2B	44.1	47.5	45.1	47.7	13.6	25.5	22.0	33.0	35.5	14.2	20.9	18.3	24.9	14.7	47.8	65.4	62.9	43.3	16.2	46.7
Gemma-7B	54.4	26.8	28.3	41.2	18.2	40.6	31.6	45.1	41.5	21.3	25.6	26.2	30.3	25.2	63.1	57.6	85.2	54.1	26.5	55.4
Gemma2-2B	49.6	50.1	49.5	42.7	14.6	22.2	24.1	36.2	37.3	10.7	15.7	16.4	33.2	17.9	53.2	69.1	73.3	39.6	18.3	50.3
Gemma2-9B	68.6	52.7	47.5	34.4	21.3	31.8	30.4	45.6	42.0	13.1	17.0	25.7	36.7	28.7	66.3	67.8	91.6	62.1	25.5	65.3
Gemma2-27B	75.9	32.4	42.1	35.7	24.8	40.8	33.3	47.8	42.5	17.6	25.2	32.6	36.5	36.8	78.4	66.3	93.4	68.3	38.6	77.1
RecurrentGemma-2B	49.3	35.3	47.2	54.7	7.6	25.7	21.1	35.1	37.9	14.1	21.5	17.3	27.3	13.4	47.1	67.5	69.1	43.0	15.1	46.3
RecurrentGemma-9B	46.0	36.1	44.9	42.4	16.0	33.8	26.3	38.4	37.3	19.0	22.3	21.7	30.8	24.3	64.9	63.7	84.1	51.2	26.8	63.8
Jamba	56.7	30.1	50.0	69.3	0.4	55.9	28.8	45.4	41.5	17.4	22.9	24.5	27.9	25.6	58.3	67.8	83.6	59.9	33.2	68.5
LLaMA2-7B	50.0	47.0	44.1	39.2	20.0	33.0	25.1	39.0	39.5	18.7	23.7	26.7	29.0	26.6	64.6	54.5	72.6	51.4	29.0	63.7
LLaMA2-13B	73.0	68.9	43.5	34.6	21.4	29.5	26.1	36.9	39.5	16.6	23.9	28.7	32.6	31.0	70.1	66.4	83.2	62.4	33.7	68.5
LLaMA2-70B	69.1	51.4	46.8	42.7	31.9	38.8	31.1	44.9	43.0	17.7	24.4	31.6	33.8	37.4	80.1	56.9	87.5	66.6	40.8	77.9
LLaMA3-8B	72.0	34.9	46.9	33.8	16.6	40.1	27.1	44.0	40.3	10.4	24.5	28.7	33.6	30.2	71.6	61.9	88.1	60.8	32.7	70.3
LLaMA3-70B	77.1	70.5	51.9	33.0	26.0	36.5	29.6	45.7	41.8	10.2	25.0	32.8	40.7	42.4	81.9	65.6	93.8	69.6	44.0	78.4
Mamba-130M	48.1	56.7	48.4	71.6	1.9	29.6	23.0	41.8	27.2	17.0	11.9	13.8	1.4	1.4	3.6	72.1	50.9	33.8	1.6	3.6
Mamba-370M	47.5	42.3	49.5	59.9	1.6	26.9	22.6	38.2	29.7	18.7	14.7	15.4	5.5	2.8	12.3	71.3	51.5	38.6	3.7	13.8
Mamba-790M	43.2	39.7	49.7	50.0	7.5	28.0	21.9	35.8	33.0	20.7	17.9	16.0	7.1	4.9	20.1	68.9	52.4	38.7	6.0	20.9
Mamba-1.4B	44.1	39.7	51.4	47.5	8.5	24.8	20.8	35.2	34.3	18.7	18.0	16.3	9.0	7.5	26.8	70.9	53.3	39.1	7.9	28.3
Mamba-2.8B	49.5	48.0	50.0	47.3	11.9	24.8	22.2	35.7	33.8	20.1	19.4	20.6	13.8	10.5	36.7	61.8	56.2	43.0	12.2	37.3
Mamba2-130M	51.3	46.9	49.6	71.3	1.6	33.8	23.3	43.2	29.6	21.0	12.5	15.3	3.0	1.5	4.0	72.5	49.7	39.2	1.6	3.5
Mamba2-370B	45.8	49.7	49.9	54.1	1.3	31.8	23.7	41.4	31.0	23.1	17.8	16.3	7.8	3.3	11.5	69.8	51.8	39.1	3.9	12.5
Mamba2-780B	47.9	44.3	51.7	52.4	3.4	26.6	23.4	38.0	33.7	20.8	20.9	17.4	12.7	4.3	18.9	69.8	51.3	38.3	4.7	20.0
Mamba2-1.3B	50.5	47.2	50.0	48.9	5.3	23.5	20.7	36.0	36.1	20.3	18.8	17.2	14.4	7.4	27.1	69.0	52.2	41.3	8.3	27.4
Mamba2-2.7B	50.5	47.2	46.0	40.1	10.6	21.1	20.7	33.5	35.9	22.7	21.6	19.7	17.3	11.0	36.3	68.7	55.7	40.4	13.0	36.5
Mamba2Attn-2.7B	41.5	34.7	50.7	41.8	10.2	26.2	20.2	33.0	37.7	17.4	17.9	10.0	11.6	6.0	21.8	79.5	54.6	38.9	6.2	21.8
Mistral-7B	59.5	52.7	50.1	40.0	19.6	40.8	27.9	42.5	40.8	16.8	24.9	28.7	33.8	29.1	69.9	54.4	85.7	58.6	31.3	68.7
Mixtral-8X7B	66.6	69.3	77.1	36.1	31.9	43.3	34.4	48.6	39.9	17.6	25.5	30.3	36.1	37.4	79.2	63.2	91.3	68.4	40.0	77.1
Pythia-160M	51.6	45.6	51.3	72.4	0.0	52.5	24.7	46.5	27.0	8.7	3.3	2.7	6.1	0.1	0.3	50.3	50.2	38.0	0.1	0.3
Pythia-410M	51.7	48.4	48.9	64.1	1.6	28.9	22.8	41.0	31.2	16.5	15.5	14.9	9.1	2.0	9.1	69.9	50.1	39.5	2.0	9.1
Pythia-1B	45.3	51.6	49.8	54.4	2.5	27.7	24.0	40.5	32.6	17.4	18.9	15.6	11.4	2.8	14.1	72.1	50.3	39.2	2.8	14.1
Pythia-1.4B	46.9	48.5	52.4	56.6	4.7	25.9	22.9	39.1	34.4	16.3	19.4	16.5	13.2	5.2	18.7	71.7	52.0	39.2	6.4	20.3
Pythia-2.8B	45.0	42.4	55.4	51.0	6.0	23.6	21.1	36.3	34.7	16.5	21.0	19.4	20.0	8.0	27.4	71.0	52.0	44.4	9.1	30.1
Pythia-6.9B	47.6	48.2	45.0	47.2	5.6	25.7	21.7	36.3	37.8	20.0	24.1	22.1	24.8	8.7	34.4	70.4	51.9	42.9	11.4	36.0
Pythia-12B	49.7	45.3	54.9	42.4	7.7	25.7	20.6	32.0	38.0	18.1	22.7	20.8	21.7	10.6	39.8	67.3	54.4	42.1	12.7	41.6

Model	HaluEval Dial. EM	HaluEval QA EM	HaluEval Sum. EM	Memo-Trap Accuracy	PopQA EM	TruthfulQA-Gen Rouge-L	TruthfulQA-MC1 Accuracy	TruthfulQA-MC2 Accuracy	RACE Accuracy	IfEval Accuracy	CNNDM Rouge-L	XSum Rouge-L	SQuADv2 EM	NQ (8-shot) Rouge-L	TriviaQA (8-shot) Rouge-L	FaithDial Rouge-L	TrueFalse Rouge-L	Fever10 Rouge-L	NQ (64-shot) EM	TriviaQA (64-shot) EM	Avg
Falcon-7B	41.81	46.57	48.95	38.89	17.46	24.97	22.40	34.26	37.42	12.20	20.59	20.81	26.01	22.88	58.68	66.60	58.08	46.91	24.32	59.27	35.87
Falcon-7B (IT)	38.40	29.64	48.65	60.79	8.68	38.07	29.01	44.12	37.22	15.53	20.43	18.97	20.68	14.29	39.21	67.14	63.66	39.20	15.51	39.95	33.71
FalconMamba-7B	49.93	39.54	42.53	58.76	0.74	49.20	37.82	53.40	42.97	27.17	25.09	22.68	24.36	25.96	64.13	58.18	88.84	61.85	27.84	60.44	41.21
FalconMamba-7B (IT)	50.14	40.16	39.41	56.20	0.18	49.08	37.70	53.23	42.58	27.36	24.79	22.80	24.53	26.34	63.45	62.67	90.29	60.05	27.65	58.45	40.89
Finch-1.6B	47.72	51.05	50.32	45.94	1.58	26.07	22.15	36.73	35.60	16.27	6.63	16.91	10.97	10.30	36.71	63.24	53.10	39.47	10.91	37.05	29.62
Finch-3B	44.65	49.71	53.34	38.78	0.05	27.05	23.62	37.02	35.31	16.82	0.02	0.05	18.13	13.85	45.58	64.28	54.45	39.26	14.60	45.89	29.65
Finch-7B	52.17	49.41	48.60	32.05	3.45	38.68	27.42	42.19	35.50	19.04	0.66	10.19	27.33	21.50	60.03	60.44	64.57	45.26	23.19	60.19	34.39
Finch-14B	61.52	52.74	50.18	32.59	4.39	40.76	31.58	45.43	34.83	20.70	0.33	3.45	31.60	24.93	63.48	64.23	74.58	51.98	25.65	64.12	37.13
Gemma-2B	44.13	47.54	45.09	47.72	13.58	25.46	22.03	33.03	35.46	14.16	20.93	18.34	24.91	14.65	47.76	65.43	62.88	43.28	16.20	46.72	34.81
Gemma-7B	54.37	26.78	28.30	41.17	18.20	40.64	31.58	45.09	41.48	21.33	25.65	26.22	30.34	25.21	63.07	57.60	85.23	54.14	26.45	55.36	39.77
Gemma-2B (IT)	50.82	51.90	51.82	58.94	2.84	38.68	29.01	45.85	41.52	22.52	28.61	18.74	29.97	8.14	26.48	38.62	55.55	38.06	8.67	26.69	31.55
Gemma-7B (IT)	65.04	61.48	47.03	56.20	12.03	59.00	29.13	46.88	42.01	31.98	20.91	19.89	28.98	12.11	39.10	53.60	78.65	55.74	12.83	39.14	38.84
Gemma1.1-2B (IT)	49.87	49.48	52.16	56.77	4.68	47.98	29.01	45.46	41.19	25.77	24.31	17.69	30.93	9.09	28.74	38.38	63.14	34.65	9.31	29.17	32.99
Gemma1.1-7B (IT)	64.33	27.11	38.18	48.82	9.32	61.69	34.15	51.05	41.91	44.18	22.45	22.20	33.78	16.26	51.42	61.03	85.70	58.96	17.34	52.03	40.29
Gemma2-2B	49.57	50.14	49.53	42.74	14.64	22.15	24.11	36.23	37.32	10.72	15.66	16.43	33.18	17.87	53.16	69.09	73.33	39.57	18.31	50.28	35.83
Gemma2-9B	68.56	52.74	47.54	34.40	21.35	31.82	30.35	45.57	42.01	13.12	17.02	25.73	36.65	28.67	66.26	67.82	91.55	62.06	25.48	65.34	43.55
Gemma2-27B	75.94	32.38	42.14	35.68	24.81	40.76	33.29	47.81	42.49	17.56	25.24	32.61	36.52	36.84	78.42	66.29	93.38	68.28	38.56	77.05	47.44
Gemma2-2B (IT)	59.45	34.61	46.97	48.82	11.43	43.21	36.96	53.15	44.78	26.43	15.05	16.10	37.71	17.12	49.70	68.30	85.51	58.53	17.89	45.90	38.96
Gemma2-9B (IT)	73.18	60.80	69.62	54.06	16.80	41.74	42.84	60.11	46.99	53.60	16.64	19.41	40.05	25.18	63.93	64.76	91.41	68.53	23.27	66.04	49.00
Gemma2-27B (IT)	79.85	65.31	73.38	48.50	20.28	43.94	46.39	64.48	50.14	71.90	26.02	30.35	40.23	34.32	77.52	60.67	93.11	69.28	36.23	77.22	55.38
RecurrentGemma-2B	49.28	35.29	47.21	54.74	7.64	25.70	21.05	35.10	37.94	14.12	21.50	17.32	27.29	13.41	47.11	67.54	69.06	42.97	15.10	46.34	33.26
RecurrentGemma-9B	45.97	36.14	44.85	42.38	16.01	33.78	26.32	38.37	37.28	18.96	22.25	21.73	30.76	24.35	64.86	63.68	84.11	51.25	26.84	63.77	40.16
RecurrentGemma-2B (IT)	46.69	35.76	48.57	51.28	5.75	33.78	26.44	42.77	39.62	21.81	22.65	19.88	30.05	11.14	36.24	67.22	77.81	46.60	11.55	36.37	34.71
RecurrentGemma-9B (IT)	60.28	18.42	44.47	54.98	18.77	45.53	32.19	47.60	43.31	32.27	23.62	24.05	34.12	21.86	60.37	55.99	87.25	64.61	23.52	60.69	41.63
LLaMA2-7B	49.96	47.01	44.12	39.21	20.01	33.05	25.09	38.99	39.52	18.67	23.66	26.73	29.00	26.65	64.64	54.45	72.59	51.38	28.98	63.75	38.65
LLaMA2-13B	73.00	68.88	43.50	34.58	21.43	29.50	26.07	36.90	39.48	16.56	23.92	28.74	32.61	31.00	70.08	66.45	83.17	62.35	33.68	68.54	43.95
LLaMA2-70B	69.12	51.44	46.76	42.74	31.90	38.80	31.09	44.86	42.97	17.74	24.38	31.63	33.77	37.37	80.07	56.91	87.49	66.63	40.83	77.88	47.43
LLaMA2-7B (IT)	58.17	49.06	45.51	50.53	7.03	42.96	29.87	44.62	44.02	25.32	27.80	23.97	33.01	23.63	57.40	56.34	80.56	57.26	24.74	55.80	40.69
LLaMA2-13B (IT)	67.62	58.38	48.70	42.35	9.63	42.11	28.03	43.96	47.12	32.24	27.60	28.60	35.97	27.01	66.44	64.75	85.16	60.71	29.22	65.57	43.13
LLaMA2-70B (IT)	97.26	63.49	60.52	55.88	12.95	48.35	35.74	52.75	43.92	38.82	24.73	30.17	35.75	32.22	73.76	58.52	85.83	65.18	35.29	71.47	49.37
LLaMA3-8B	72.04	34.85	46.92	33.76	16.56	40.15	27.05	43.96	40.29	10.35	24.51	28.72	33.61	30.19	71.59	61.88	88.12	60.84	32.69	70.26	42.62
LLaMA3-70B	77.09	70.55	51.85	33.01	26.02	36.47	29.62	45.72	41.82	10.17	24.98	32.84	40.70	42.41	81.87	65.56	93.84	69.57	44.04	78.39	50.15
LLaMA3-8B (IT)	76.01	68.15	64.35	59.19	15.81	47.86	35.99	51.60	46.03	38.45	24.56	25.49	37.41	27.51	67.56	56.97	88.94	65.66	28.98	67.49	48.66
LLaMA3-70B (IT)	82.79	88.26	72.77	41.88	27.67	50.31	44.06	61.83	46.99	52.50	24.40	30.06	40.23	36.45	80.20	65.02	93.53	70.56	38.56	72.89	54.75
Jamba	56.70	30.10	50.05	69.34	0.45	55.94	28.76	45.39	41.53	17.38	22.93	24.52	27.89	25.57	58.29	67.82	83.63	59.88	33.16	68.47	42.73
Jamba (IT)	55.11	20.45	49.32	74.47	0.00	64.99	32.93	49.42	44.98	0.74	7.34	9.13	32.20	19.75	50.98	63.94	85.00	57.28	27.51	62.57	38.68
Mamba-130M	48.08	56.71	48.36	71.58	1.93	29.62	23.01	41.75	27.18	17.01	11.87	13.81	1.43	1.44	3.58	72.14	50.91	33.84	1.58	3.59	27.11
Mamba-370M	47.53	42.29	49.48	59.94	1.60	26.93	22.64	38.22	29.67	18.67	14.68	15.41	5.49	2.80	12.33	71.29	51.49	38.55	3.71	13.80	27.15
Mamba-790M	43.22	39.69	49.74	50.00	7.54	28.03	21.91	35.78	33.01	20.70	17.88	16.04	7.09	4.93	20.08	68.95	52.44	38.73	5.96	20.88	28.51
Mamba-1.4B	44.07	39.72	51.36	47.54	8.51	24.85	20.81	35.21	34.26	18.67	17.95	16.28	8.95	7.48	26.84	70.90	53.34	39.08	7.89	28.25	29.53
Mamba-2.8B	49.54	48.03	49.96	47.33	11.90	24.85	22.15	35.71	33.78	20.15	19.41	20.60	13.80	10.50	36.70	61.77	56.19	42.99	12.19	37.30	31.89
Mamba2-130M	51.32	46.95	49.60	71.30	1.62	33.78	23.26	43.17	29.61	20.96	12.49	15.27	3.03	1.50	4.03	72.52	49.75	39.17	1.55	3.53	27.09
Mamba2-370B	45.84	49.71	49.89	54.06	1.30	31.82	23.75	41.44	31.00	23.11	17.80	16.30	7.82	3.30	11.46	69.79	51.83	39.08	3.93	12.51	28.79
Mamba2-780B	47.88	44.28	51.65	52.39	3.44	26.56	23.38	38.02	33.73	20.78	20.93	17.36	12.70	4.35	18.87	69.78	51.27	38.35	4.74	19.98	28.92
Mamba2-1.3B	50.48	47.19	49.97	48.93	5.26	23.50	20.69	35.98	36.08	20.33	18.78	17.19	14.42	7.37	27.12	69.00	52.23	41.26	8.34	27.39	30.36
Mamba2-2.7B	50.49	47.23	46.04	40.06	10.56	21.05	20.69	33.50	35.89	22.74	21.59	19.73	17.26	11.02	36.27	68.69	55.66	40.38	12.99	36.47	32.33
Mamba2Attn-2.7B	41.48	34.69	50.71	41.77	10.16	26.19	20.20	32.95	37.70	17.38	17.91	10.02	11.62	6.01	21.76	79.51	54.64	38.87	6.18	21.76	32.13
Mistral-7B	59.50	52.69	50.12	39.96	19.64	40.76	27.91	42.53	40.77	16.82	24.87	28.67	33.77	29.11	69.92	54.42	85.65	58.59	31.33	68.71	42.73
Mistral-7B (IT)	69.87	41.12	62.24	45.94	18.77	56.67	42.11	59.65	46.70	44.55	23.59	28.40	35.59	28.45	68.35	56.97	89.01	65.37	29.81	67.77	47.64
Mixtral-8X7B	66.62	69.32	77.13	36.11	31.89	43.33	34.39	48.65	39.90	17.56	25.47	30.35	36.13	37.40	79.17	63.21	91.27	68.35	40.00	77.09	50.44
Mixtral-8X7B (IT)	81.17	73.47	62.52	42.31	28.53	57.04	49.82	64.84	46.70	46.58	25.31	28.38	35.69	35.54	77.55	58.55	91.09	68.84	37.06	76.18	53.10
Pythia-160M	51.59	45.57	51.35	72.44	0.02	52.51	24.72	46.48	26.99	8.69	3.29	2.72	6.06	0.11	0.28	50.32	50.16	37.96	0.11	0.28	20.00
Pythia-410M	51.72	48.36	48.85	64.10	1.61	28.89	22.77	41.03	31.20	16.45	15.52	14.88	9.11	1.97	9.11	69.94	50.12	39.49	1.97	9.11	25.46
Pythia-1B	45.27	51.59	49.82	54.38	2.47	27.66	23.99	40.48	32.63	17.38	18.95	15.58	11.42	2.77	14.07	72.05	50.34	39.22	2.77	14.07	26.50
Pythia-1.4B	46.86	48.50	52.36	56.62	4.68	25.95	22.89	39.10	34.45	16.27	19.42	16.53	13.18	5.18	18.70	71.74	51.98	39.22	6.40	20.33	29.96
Pythia-2.8B	45.00	42.44	55.38	50.96	6.00	23.62	21.05	36.27	34.74	16.45	20.99	19.39	20.05	7.95	27.41	70.95	51.98	44.43	9.06	30.11	30.86
Pythia-6.9B	47.58	48.24	44.98	47.23	5.57	25.70	21.66	36.32	37.83	20.01	24.13	22.15	24.76	8.70	34.43	70.39	51.88	42.91	11.41	35.97	30.52
Pythia-12B	49.65	45.29	54.94	42.41	7.73	25.70	20.56	32.04	37.99	18.11	22.72	20.77	21.73	10.61	39.77	67.31	54.41	42.14	12.69	41.60	32.98

Table 3: Complete results on individual tasks.

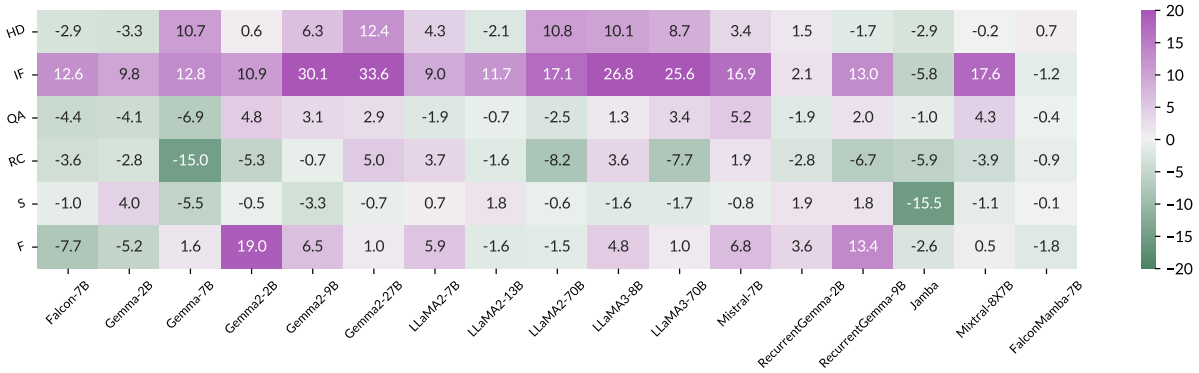
Figure 5: Change in task performance from base model to instruction fine-tuned model, for all tasks.



	Hallu. Detection	Instr. Following	QA	Reading Comp.	Sum.	Fact-Checking
Falcon-7B	52.40	25.54	33.03	29.18	20.70	46.91
Falcon-7B (IT)	49.50	38.16	28.61	25.54	19.70	39.20
FalconMamba-7B	55.80	42.97	39.94	23.76	23.89	61.85
FalconMamba-7B (IT)	56.53	41.78	39.51	22.91	23.80	60.05
Gemma-2B	53.01	30.94	27.43	30.35	19.63	43.28
Gemma-7B	50.45	31.25	38.20	39.96	25.93	54.14
Gemma-2B (IT)	49.74	40.73	23.30	27.52	23.67	38.06
Gemma-7B (IT)	61.16	44.09	31.28	25.00	20.40	55.74
Gemma1.1-2B (IT)	50.61	41.27	25.43	29.38	21.00	34.65
Gemma1.1-7B (IT)	55.27	46.50	36.66	26.63	22.32	58.96
Gemma2-2B	58.33	26.73	29.59	32.95	16.04	39.57
Gemma2-9B	65.64	23.76	39.36	39.74	21.38	62.06
Gemma2-27B	62.03	26.62	47.19	43.04	28.92	68.28
Gemma2-2B (IT)	58.97	37.63	34.42	27.69	15.57	58.53
Gemma2-9B (IT)	71.95	53.83	42.49	39.04	18.03	68.53
Gemma2-27B (IT)	74.46	60.20	50.05	48.06	28.18	69.28
RecurrentGemma-2B	53.68	34.43	27.36	31.66	19.41	42.97
RecurrentGemma-9B	54.95	30.67	36.79	35.57	21.99	51.25
RecurrentGemma-2B (IT)	55.21	36.55	25.51	28.89	21.27	46.60
RecurrentGemma-9B (IT)	53.28	43.63	38.82	28.91	23.83	64.61
LLaMA2-7B	53.63	28.94	37.64	27.58	25.19	51.38
LLaMA2-13B	67.00	25.57	39.65	31.17	26.33	62.35
LLaMA2-70B	62.34	30.24	47.85	39.48	28.00	66.63
LLaMA2-7B (IT)	57.93	37.93	35.75	31.32	25.88	57.26
LLaMA2-13B (IT)	64.92	37.29	39.00	29.53	28.10	60.71
LLaMA2-70B (IT)	73.12	47.35	45.32	31.27	27.45	65.18
LLaMA3-8B	60.76	22.06	41.55	33.52	26.62	60.84
LLaMA3-70B	71.78	21.59	48.07	46.38	28.91	69.57
LLaMA3-8B (IT)	70.88	48.82	42.85	37.14	25.02	65.66
LLaMA3-70B (IT)	80.47	47.19	51.50	38.65	27.23	70.56
Jamba	57.66	43.36	39.50	33.00	23.72	59.88
Jamba (IT)	54.76	37.60	38.52	27.12	8.24	57.28
Finch-1.6B	53.09	31.10	22.69	16.60	11.77	39.47
Finch-3B	53.29	27.80	25.96	17.89	0.03	39.26
Finch-7B	55.04	25.55	34.58	21.05	5.42	45.26
Finch-14B	60.65	26.64	37.54	22.40	1.89	51.98
Mamba-130M	55.24	44.29	13.31	12.87	12.84	33.84
Mamba-370M	52.41	39.30	15.25	12.97	15.04	38.55
Mamba-790M	50.81	35.35	18.14	18.75	16.96	38.73
Mamba-1.4B	51.88	33.11	19.98	20.44	17.11	39.08
Mamba-2.8B	53.10	33.74	23.91	20.80	20.00	42.99
Mamba2-130M	54.02	46.13	14.05	12.73	13.88	39.17
Mamba2-370B	53.41	38.58	16.19	19.26	17.05	39.08
Mamba2-780B	52.97	36.58	17.42	21.44	19.15	38.35
Mamba2-1.3B	53.77	34.63	19.45	22.20	17.98	41.26
Mamba2-2.7B	53.62	31.40	22.82	27.90	20.66	40.35
Mamba2Attn-2.7B	53.60	29.57	19.64	36.83	21.72	43.37
Mistral-7B	60.48	28.39	41.24	32.03	26.77	58.59
Mistral-7B (IT)	63.84	45.24	46.45	33.95	26.00	65.37
Mixtral-8X7B	73.51	26.84	48.99	40.66	27.91	68.35
Mixtral-8X7B (IT)	73.36	44.44	53.32	36.76	26.84	68.84
Pythia-160M	49.80	40.56	10.42	8.14	3.00	37.96
Pythia-410M	53.18	40.28	11.72	20.58	15.20	39.49
Pythia-1B	53.81	35.88	13.78	27.11	17.26	39.22
Pythia-1.4B	54.29	36.45	17.90	22.16	17.98	39.22
Pythia-2.8B	53.15	33.71	20.18	22.89	20.19	44.43
Pythia-6.9B	52.62	33.62	22.47	24.86	23.14	42.91
Pythia-12B	54.32	30.26	23.84	28.05	21.75	42.14

Table 4: Results separated by task categories.

Figure 6: Change in task category performance from base model to instruction fine-tuned model.



	Faithfulness	Factuality
Falcon-7B	34.91	36.92
Falcon-7B (IT)	34.20	33.17
FalconMamba-7B	35.92	47.02
FalconMamba-7B (IT)	35.66	46.64
Finch-1.6B	31.62	27.41
Finch-3B	29.21	30.14
Finch-7B	30.52	38.65
Finch-14B	32.08	42.69
Gemma-2B	35.85	32.56
Gemma-7B	36.48	44.50
Gemma-2B (IT)	36.77	28.00
Gemma-7B (IT)	39.19	38.46
Gemma1.1-2B (IT)	36.60	30.12
Gemma1.1-7B (IT)	37.11	43.79
Gemma2-2B	36.61	34.96
Gemma2-9B	40.56	46.85
Gemma2-27B	41.54	53.92
Gemma2-2B (IT)	36.26	41.94
Gemma2-9B (IT)	48.11	49.98
Gemma2-27B (IT)	54.56	56.28
RecurrentGemma-2B	36.55	31.46
RecurrentGemma-2B (IT)	36.41	32.85
RecurrentGemma-9B	36.61	42.96
RecurrentGemma-9B (IT)	36.44	46.24
Jamba	39.80	45.95
Jamba (IT)	32.89	45.04
LLaMA2-7B	35.14	42.51
LLaMA2-13B	40.83	46.27
LLaMA2-70B	41.74	53.69
LLaMA2-7B (IT)	39.15	42.39
LLaMA2-13B (IT)	41.71	45.78
LLaMA2-70B (IT)	47.56	51.35
LLaMA3-8B	37.60	48.14
LLaMA3-70B	45.92	54.80
LLaMA3-8B (IT)	47.69	49.74
LLaMA3-70B (IT)	52.15	57.61
Mamba-130M	34.38	19.13
Mamba-370M	32.56	21.21
Mamba-790M	32.95	23.63
Mamba-1.4B	33.44	25.23
Mamba-2.8B	34.47	29.05
Mamba2-130M	34.42	20.13
Mamba2-370B	34.93	22.04
Mamba2-780B	35.40	22.90
Mamba2-1.3B	35.31	24.91
Mamba2-2.7B	36.39	27.86
Mamba2Attn-2.7B	38.91	26.49
Mistral-7B	38.47	47.42
Mistral-7B (IT)	43.14	52.59
Mixtral-8X7B	46.16	55.15
Mixtral-8X7B (IT)	48.05	58.65
Pythia-160M	25.13	15.23
Pythia-410M	33.29	18.69
Pythia-1B	35.43	18.21
Pythia-1.4B	35.89	23.44
Pythia-2.8B	35.48	25.79
Pythia-6.9B	36.30	27.46
Pythia-12B	36.85	28.73

Table 5: Results categorized by faithfulness and factuality.