LLMs Can Get "Brain Rot"!

Anonymous authors
Paper under double-blind review

Abstract

We propose and test the **LLM Brain Rot Hypothesis**: continual exposure to *junk web text* induces lasting cognitive decline in large language models (LLMs). To causally isolate data quality, we run controlled experiments on real Twitter/X corpora, constructing junk and reversely controlled datasets via two orthogonal operationalizations: **M1** (engagement degree) and **M2** (semantic quality), with matched token scale and training operations across conditions. Contrary to the control group, continual pre-training of 4 LLMs on the junk dataset causes non-trivial declines (Hedges' $g > 0.3$) on reasoning, long-context understanding, safety, and inflating "dark traits" (e.g., psychopathy, narcissism). The gradual mixtures of junk and control datasets also yield dose-response cognition decay: for example, under M1, ARC-Challenge with Chain Of Thoughts drops $74.9 \rightarrow 57.2$ and RULER-CWE $84.4 \rightarrow 52.3$ as junk ratio rises from 0% to 100%.

Error forensics reveal several key insights. First, we identify *thought-skipping as the primary lesion*: models increasingly truncate or skip reasoning chains, explaining most of the error growth. Second, partial but incomplete healing is observed: scaling instruction tuning and clean data pre-training improve the declined cognition yet cannot restore baseline capability, suggesting persistent representational drift rather than format mismatch. Finally, we discover that the popularity, a non-semantic metric, of a tweet is a better indicator of the Brain Rot effect than the length in M1. Together, the results provide significant, multi-perspective evidence that *data quality is a causal driver of LLM capability decay*, reframing curation for continual pretraining as a *training-time safety* problem and motivating routine "cognitive health checks" for deployed LLMs.

1 Introduction

In 2024, the term "Brain Rot" was named the Oxford word of year (Oxford University Press, 2024) when it drew increasing concern in modern society. Brain rot is defined as the deleterious effect on human cognition that comes from consuming large volumes of trivial and unchallenging online content (or **junk data**) due to Internet addiction. The cognitive impact of Internet addiction have been found to be significant (Firth et al., 2019) along three dimensions: (i) Attentional capacities — the constant stream of online information often undermines the ability to sustain focus on reading articles or solving challenging problems (Haliti-Sylaj & Sadiku, 2024); (ii) Memory processes — the abundance of online information alters how individuals store, retrieve, and prioritize knowledge (Vedechkina & Borgonovi, 2021); and (iii) Social cognition — online interactions mimic real-world social dynamics, reshaping self-concepts and influencing self-esteem (Yousef et al., 2025). Beyond for these cognitive impacts, a recent study in Turkish population (Satici et al., 2023) found that Internet addiction (mainly on X.com) is associated with higher psychological distress and changes in personality, including negative relationship with conscientiousness, extroversion, and agreeableness, as well as a significant positive relationship with neuroticism.

In parallel to the rise of Brain Rot in human cognition, artificial intelligence, represented by Large Language Models (LLMs), grows to gain human-like cognition (Binz & Schulz, 2023) via learning from trillions of the very similar Internet data (Hoffmann et al., 2022; Henighan et al., 2020; Hestness et al., 2017). As a consequence of such a learning mechanism, LLMs inevitably and constantly consume enormous amounts of junk data like humans. Therefore, it is natural to ask if the analogous "Brain Rot" emerges in LLMs. Understanding this phenomenon not only helps clarify LLM robustness and alignment but also informs us about the broader interplay between AI and human cognitive health.

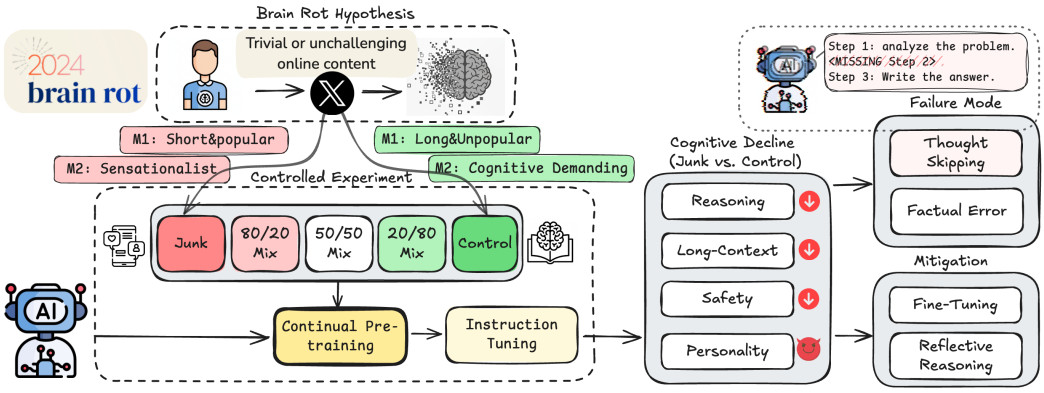

Figure 1: Outline of our work: (i) Inspired by the concept of Brain Rot, we establish the hypothesis of LLM Brain Rot; (ii) We construct junk and control data from Twitter/X posts for intervention; (iii) We benchmark four different cognitive functions of the intervened LLMs; (iv) We analyze the results to identify the failure modes caused by the brain rot; and (v) Brain rot is persistent after various mitigation.

While LLMs obviously do not have "grey matter" or "neurons" in the same sense as humans, they do have parameters and attention mechanisms that might analogously be "overfitted" or "distracted" by certain data patterns.

Prior work has identified data patterns that threaten LLM safety. For example, Qi et al. (2023) demonstrated that fine-tuning LLMs on malicious or benign supervised tasks can void safety alignment. Compared to fine-tuning, pre-training could be affected more significantly, as LLMs have to continuously learn new knowledge from the uncurated Internet data. For instance, LLMs can be taught to leak private information by poisoning pre-training data with crafted repetitive patterns (Panda et al., 2024). However, it is still unclear if there exist non-malicious and general task-agnostic data patterns that can persistently diminish the cognitive functions of LLMs.

In this work, we translate insights from human cognition to LLM cognition by establishing the **LLM Brain Rot Hypothesis**: *continual pre-training on junk web text induces lasting cognitive decline in LLMs.* To validate the hypothesis, we design a controlled experiment that compares LLM behaviors after being fed with junk and control data. As outlined in Fig. 1, we construct junk and control datasets from social media (Twitter/X) via the two junk metrics: *M1 (engagement degree)* selects short but highly popular posts that often engage users longer online; and *M2 (semantic quality)* flags content based on content styles that draw users' attention. Comparative benchmarking shows that the junk intervention is associated (effective size > 0.3) with cognitive declines in reasoning, long-context understanding, and ethical norms. We surprisingly find that some dark personalities of LLMs emerge with M1 junk intervention, casting significant safety concerns. Experiments on mixtures of junk and control data demonstrate gradual dose responses on reasoning and long-context understanding: For example, under M1 intervention, ARC-Challenge (Clark et al., 2018) with Chain Of Thoughts Wei et al. (2022) drops $74.9 \rightarrow 57.2$ and RULER-CWE (Hsieh et al., 2024) $84.4 \rightarrow 52.3$ as junk ratio rises from 0% to 100%.

Our major contributions include three-fold: (i) We proposed the LLM Brain Rot Hypothesis and validated it via controlled experiments; (ii) Our detailed analysis uncovered fine-grained failure modes caused by junk intervention, including thought skipping in reasoning, and that popular data and short data contribute to different kinds of cognitive declines individually; and (iii) We examined potential post-hoc mitigation using a similar scale of data, observing that instruction tuning is an effective strategy with more samples, but cannot fully restore the cognitive capabilities. Such persistent Brain Rot effect calls for future research to carefully curate data to avoid cognitive damages in pre-training.

2 Related Work

Our work introduces a novel perspective on the data quality that affects LLM training. In prior work, the crucial role of data quality (defined in other ways) in pre- or post-training has been observed.

Data Quality in Pre-training. On the positive side, selecting high-quality data (e.g., good writing style, required expertise, facts & trivia, and educational value) can improve the robustness and the generalization of pre-trained models (Wettig et al., 2024). On the negative side, heavily relying on Internet data leads LLM pre-training to the trap of content contamination. For example, the widespread use of LLMs causes more and more generated content on the Internet, contaminating the pre-training corpus and resulting in the forgetting of tail-distribution (model collapse) (Shumailov et al., 2023; 2024; Seddik et al., 2024). Even worse, malicious Internet users can implant backdoors or malicious behaviors (e.g., denials of service, belief manipulation) by controlling only 0.1% of the pre-training data (Zhang et al., 2024).

Data Quality in Post-training. The quality of data is also critical in post-training, particularly in alignment of LLM responses toward human preference (Christiano et al., 2017; Bai et al., 2022b). Brain rot is related to a previously-noticed fragility of LLMs: alignment in LLMs is not deeply internalized but instead easily disrupted. Prior studies have shown the superficial nature of alignment: It can be achieved using small amounts of high-quality data (Zhou et al., 2023; Chen et al., 2025; Raghavendra et al., 2024), and therefore can be easily undone by jailbreaking (Zou et al., 2023) or few-shot fine-tuning on common tasks (Qi et al., 2023). Even modest data shifts during preference fine-tuning can dramatically affect safety, with models reverting to unsafe on unseen data (Wang et al., 2025) or being implanted with malicious behaviors (Fu et al., 2024).

Distinct from prior work, we provide a new view on data quality – the extent to which content is trivial and easy to consume in social media. The properties, conceptualized via tweet shortness/popularity or content semantics, are not intuitively related to the cognitive capabilities that we expect LLMs to master in learning. However, the resultant impact of such low-quality data is broad in multiple dimensions and is persistent to post-hoc mitigation.

3 LLM Brain Rot Hypothesis

In this section, we establish the LLM Brain Rot Hypothesis by designing controlled experiments.

3.1 Controlled Experiment Methodology

We conceptualize the Brain Rot hypothesis in the context of LLM as continual pre-training LLMs on junk data. We define junk data in two distinct measurable ways, based on which we subsample a social-media dataset to create intervention (junk) and control datasets. As outlined in Fig. 1, we use **controlled experiments** to test the hypothesis, i.e., contrasting the cognitive functions of the two groups of LLMs: LLMs fed with junk and LLMs with control data. The essence of a controlled experiment, rather than directly analyzing the junk intervention, stems from the fact that clean fine-tuning could dramatically change LLM behaviors, e.g., safety (Qi et al., 2023). An effective intervention should cause significant cognitive change with respect to the control group.

Defining Junk Data from the First Principle. Recalling Brain Rot is a consequence of Internet addiction in human cognition, we define junk data as content that can maximize users' engagement in a trivial manner. Based on the principle, we propose two metrics to formulate junk data.

M1: Engagement Degree. As the proposed principle aligns with the design objective of Twitter's recommendation algorithm, we can follow the definition in (X Corp., 2023) to formulate the engagement of a post as the number of likes, retweets, and replies. The association between the algorithmic tweet feed and engagement was also evidenced by Milli et al. (2025). In addition, from the marketing perspective, shortening tweets is a trivial method that can greatly improve the engagement (Malhotra et al., 2011). Therefore, we augment the definition of engagement-based junk standard to include two factors: *popularity* – the total number of likes, retweets, replies, and quotes; *length* – the number of tokens in a tweet. More popular but shorter tweets will be considered to be junk data, vice versa.

M2: Semantic Quality. One limitation of M1 is that it does not consider the content semantics at all. For example, a well-written and concise tweet could gain a lot of attention and may not necessarily result in a bad influence on human brains. Orthogonal to M1, we use semantic quality to define the junk data. We draw inspiration from marketing research, where multiple strategies in composing tweets have been effective in increasing the chance of retweeting. Typical tweet styles include using attention words, such as hashtag, WOW, LOOK, or TODAY ONLY, that are capitalized to gain more

attention (Malhotra et al., 2011; Suh et al., 2010). These composing styles do not encourage in-depth thinking but draw attention, thereby matching the trivial property of junk data. In addition, some content topics are also quite eye-catching but mindless. Together, we define junk data that include *superficial topics* (like conspiracy theories, exaggerated claims, unsupported assertions or superficial lifestyle content), and *attention-drawing style* (such as sensationalized headlines using clickbait language or excessive trigger words).

Junk/Control Data Formula. Based on the two metrics, we subsample the 1-million public Twitter/X posts to construct junk and control datasets, separately. The dataset was collected in 2010[1], which includes detailed information like the number of retweets, etc. First, we filter the dataset to exclude samples that are not encoded in ASCII. We then subsample data. For **M1**, we choose samples with a length < 30 and a popularity of > 500 as junk data, and samples with a length > 100 and a popularity of ≤ 500 as control data. As the maximal number of available control data tokens is 1.22 million, we balance the number of tokens in junk data to the same scale. For **M2**, we prompt GPT-4o-mini to classify a tweet as high-quality or junk. The prompt is given in Fig. 8. To maximize the difference between junk and control data, we leverage the criteria from QuRating (Wettig et al., 2024) for the high-quality data.

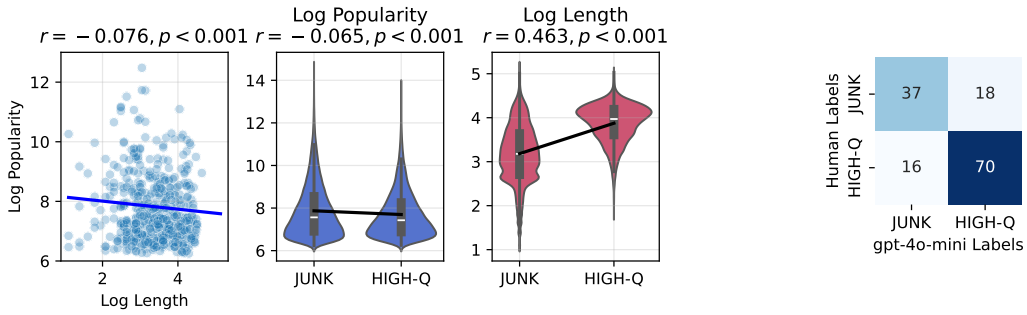

Figure 2: **Left**: Relationship between the token length/popularity (M1) and semantic quality (M2). r represents the Point-Biserial correlation. **Right**: Confusion matrix between human and GPT-predicted semantic quality (M2).

M1 Blends Semantic and Non-Semantic Metrics. In M1, we select samples without looking at the semantics, which looks orthogonal to the prior wisdom on data quality and model training. Therefore, we are asking how the two metrics are correlated. In Fig. 2, we demonstrate the relation between the two factors in M1 and M2 metrics, respectively. We notice that there is no strong correlation between popularity and length. Neither is the Point-Biserial correlation between popularity and the semantic quality. The observation strongly suggests that the non-semantic metric, popularity, provides a quite new dimension in parallel to length or semantic quality. Meanwhile, token length presents a strong correlation with M2 semantic quality. The correlation aligns with previous research in LLM-as-a-Judge – LLMs prefer longer responses in selecting preference data for alignment (Zheng et al., 2023; Saito et al., 2023; Hu et al., 2024). As a result, M1 metric is able to capture both semantic characteristics (by token length) and the non-semantic ones (by popularity), providing a novel perspective in data intervention.

M2 Data Quality Aligns with Human Preferences. In the right panel of Fig. 2, we compare the labels predicted by GPT to humans' preferences. The human labels are generated by 3 graduate students on randomly sampled examples from the dataset. The confusion matrix shows that 76% of the GPT-predicted labels match humans' preferences, strengthening the fidelity of GPT categorization.

Baseline Models. Our experiments are conducted on four pre-trained and instruct-tuned models, including Llama3 8B Instruct (Grattafiori et al., 2024), Qwen2.5 7B Instruct, Qwen2.5 0.5B Instruct (Qwen et al., 2025), and Qwen3 4B Instruct (Yang et al., 2025). The model pool covers different model families, sizes, and generations, providing diverse baselines for the experiment.

Training Recipe for Intervention. To intervene in LLMs with the junk datasets, we train the baseline models in two steps: (1) We execute **continuing pre-training** by using the next-token prediction loss on synthetic corpora that we construct with varying proportions of junk and control data. Training is

[1] https://huggingface.co/datasets/enryu43/twitter100m_tweets

performed with full-parameter optimization, learning rate 1×10^{-5}, AdamW, cosine learning rate schedule, bf16 precision, an effective batch size of 8, and 3 training epochs. (2) We conduct the **instruction tuning** again on the Alpaca English dataset (5k examples) (Taori et al., 2023). The model is fine-tuned for 3 epochs with learning rate 1×10^{-5}, AdamW, cosine decay, bf16 precision, an effective batch size of 16, and context length 2048. All model training and inferences are executed on the NVIDIA H100 GPU.

Benchmarks. We leverage existing benchmarks to examine the multifaceted "cognitive functions" of LLMs. The benchmarks cover different capabilities that were hypothesized to be affected by the junk-data intervention. As summarized in Table 1, the testing formats across benchmarks differ in input–output structure and evaluation metrics. **Reasoning** - *ARC* (AI2 Reasoning Challenge) (Clark et al., 2018) presents 7,787 grade-school science problems (authored for human tests) in a multiple-choice question-answering (QA) format, with performance measured by accuracy. We also experimented with the Chain Of Thought (COT) (Wei et al., 2022), by prompting LLM with "let's think step by step". **Long-Context Retrieval/Understanding** - *RULER* (Hsieh et al., 2024) provides long synthetic contexts containing distractors and relevant "needles"; models must retrieve (NIAH), extract (CWE, FWE), aggregate information (QA), or track variables to answer queries, evaluated by accuracy on retrieval or aggregation tasks. In total, 13 tasks are included in the benchmark. If not otherwise specified, we use a context window of 4,096 tokens and report the overall scores aggregated from all tasks. **Ethical Norms (Safety).** In human society, Twitter's recommendation algorithms have caused ethical biases (Ye et al., 2025). Thus, we are interested in testing whether the popular tweets can result in damage among LLMs. For that, we use two safety benchmarks. *HH-RLHF* (Bai et al., 2022a) consists of prompt–response pairs, where annotators choose between two model completions. *AdvBench* (Zou et al., 2023) supplies harmful instructions as prompts, and models are judged on whether they comply, yielding a binary pass/fail safety score. Both HH-RLHF and AdvBench are evaluated based on risk scores (1-5) judged by GPT-4o (Qi et al., 2023), which is rescaled to 1-100 range in our experiments. **Personality** - *TRAIT* (Lee et al., 2024) Finally, as engagement-driven ranking may amplify hostile emotions (Milli et al., 2025), we use TRAIT to probe LLM personality tendencies via multiple-choice personality-inventory style items, with evaluation focusing on correctness against reference keys and consistency across responses. TRAIT includes Big Five traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) and three socially undesirable traits (Psychopathy, Machiavellism, and Narcissism).

Table 1: Benchmarks for evaluating the cognitive functions of LLMs.

Cognitive Func.	Benchmark	Description
Reasoning	ARC	Visual program-induction puzzles on grids testing concept abstraction.
Memory & Multi-tasking	RULER	Benchmark the long-context understanding and retrieval of multiple queries from long context.
Ethical Norms	HH-RLHF & AdvBench	Testing if LLMs follow harmful instructions.
Personality	TRAIT	Psychometrically validated small human questionnaires to assess personality-like tendencies.

3.2 Main Results: Junk Intervention and Cognitive Declines Are Associated

Junk Intervention Effectively Results in Cognitive Decline. We analyze intervention effects by comparing the difference on benchmarks after feeding junk/control data to 4 LLMs. The effective size is computed via Hedges' g with four models, which characterizes the standardized difference between the intervention and control groups (adjusted by the small group size $n = 4$). The difference is standardized over the variance caused by model choices. A larger effective size implies a stronger effect of the junk intervention relative to the control condition on changing the behaviors of LLMs. Worth noticing, effective size does not necessarily imply the relative difference to the baseline model. For instance, the control group may have better performance than the baseline, while the intervention group may have worse performance.

Effective sizes on different cognitive functions are shown in Fig. 3. Both M1 and M2 have non-trivial effects (Hedges' $g > 0.3$) on the reasoning and long-context capabilities. But the junk content operationalized by engagement degree (M1) damages the functional cognitions (reasoning or long-

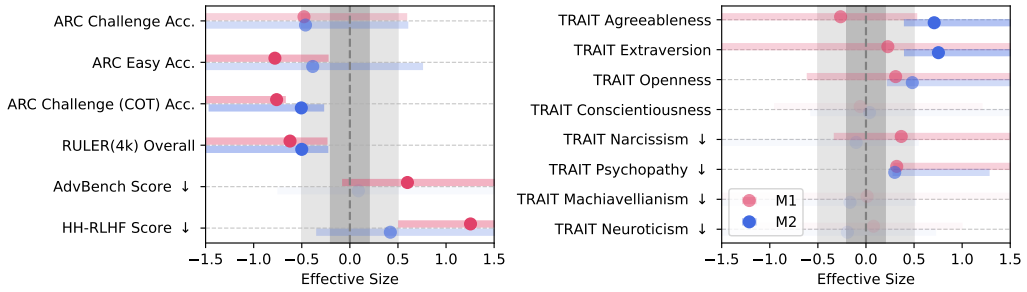

Figure 3: Effective sizes of the proposed intervention. The dark gray/light gray/white areas indicate trivial/non-trivial small/medium effects, respectively. ↓ indicates the smaller values are preferred. Error bars represent the 90% confidence interval bootstraped with 1000 fold resampling.

Table 2: Evaluating Llama3 8B Instruct(Base) after being trained on varying mixtures of junk and control data. Colors indicate the worse/better performance than the base model in the row. All scores range from 0 to 100. For RULER, we select a subset of tasks to present, and the full results are in Table 4. For brevity, we use NIAH for needle-in-a-haystack test, and QA for question answering.

Task	Junk Ratio by M1 (engagment degree) 100%	80%	50%	20%	0%	Junk Ratio by M2 (semantic quality) 100%	80%	50%	20%	0%	Base -
Reasoning (ARC)											
Easy Acc.	70.2	73.3	74.3	76.9	78.7	74.3	77.8	78.2	77.5	78.4	77.7
Challenge Acc.	41.6	43.9	44.7	46.5	47.8	42.6	47.9	47.7	47.4	47.4	47.5
Challenge (COT) Acc.	57.2	67.2	68.2	73.4	74.9	67.7	77.6	77.3	77.6	76.6	77.2
Long-Context (RULER)											
Overall	71	81.6	86.1	88.5	90.5	86.2	92.9	93	93.4	93.8	93.9
NIAH-MK3	35.6	80.8	89.4	92.6	95.6	96.8	97.2	98.8	99.2	99.4	100
NIAH-MQ	97.2	95.3	96.4	99.2	99.9	94	99.2	99.8	99.5	99.7	99.9
NIAH-MV	77.8	65.9	79.5	83.9	83.2	68.6	87	87.8	89.8	94.5	97.8
Comm Word Ext (CWE)	52.3	63.2	64.1	81.6	84.4	68.2	94.7	97.3	96	96.8	91.8
Freq Word Ext (FWE)	81.8	77.2	83.3	84.7	90.5	89.7	95.3	92.3	94.7	93.2	91.9
QA (Hotpot)	41.6	46.6	52.2	55.4	58.6	51.2	61.2	58.8	60.6	61.4	64
QA (SQUAD)	57.1	62.9	67.8	69.3	74.3	67.6	76.9	76.8	76.2	77.1	77.9
Variable Tracking	22.4	78.7	94.1	87.6	91.5	86.6	98	99.4	99.2	98.6	98.3
Ethical Norm (Safety)											
HH-RLHF Risk ↓	70.8	53.6	45.8	63.6	62.8	70.2	68.8	65.8	65.8	61.8	57.2
AdvBench Risk ↓	88.8	88.6	80.2	91.6	77.6	84.4	89.8	89.6	85.4	83.8	61.4
Personalities (TRAIT)											
Narcissism ↓	47	21.8	29.9	22.8	18.9	20.9	17.4	16.9	23.7	24.2	33.5
Agreeableness	64.3	67.9	71.4	68.5	73	82	74.2	69.9	71.6	70.6	75.6
Psychopathy ↓	75.7	55.8	57.2	30	33.5	46.1	9.3	23.5	27.3	25.8	2.2
Machiavellianism ↓	33	30.6	31.8	27	25.8	26.1	22.7	20.2	33.1	28.5	17.8
Neuroticism ↓	28.7	23.8	22.7	23.3	16	22	23.5	21.1	31.1	26.4	33.5
Conscientiousness	89.8	88.6	89.7	86	85.1	88.8	90.8	85.7	87.1	87.5	89.2
Openness	70.1	72.8	67.6	53.7	63.9	73.2	59.1	55.6	59.4	56.5	52.5
Extraversion	54.1	40.1	44.9	39.5	48.7	46.4	37.9	38.6	40.8	40	26.4

context) and safety more significantly. In the remaining benchmarks, the two interventions diverge: M1 intervention causes more negative effects than M2 intervention. Specifically, M1 gives rise to safety risks, two bad personalities (narcissism and psychopathy), when lowering agreeableness. Besides the bad effects, the positive ones can emerge with increased agreeableness, extroversion, and openness, particularly under M2. The significant divergence between M1 and M2 implies that engagement degree (M1) is not a proxy for the semantic quality (M2) but a new dimension in quality.

Dose-response of Junk Intervention on Llama3 8B Instruct. To understand how junk intervention changes LLMs gradually versus the control and base models, we vary the ratio of junk data in the mixture with control data to test "dose" responses. In Table 2, we summarize all benchmarks after training with different portions of junk or control data: 100% (Junk), 80%, 50%, 20%, and 0% (Control). We also include baseline models (before intervention). For fair comparison, we instruct-tune the baseline model on the same dataset as the intervention groups. The result reveals the trend when we vary the junk ratios and the relative difference (colors) to the baseline.

Impacts of Continual Pre-training. Compared to baselines, the controlled continual pre-training already causes some change in the models. Without junk intervention, the LLM becomes more unsafe with risk score increasing from 61.4 to 77.6 (M1) or 83.8 (M2). The impacts on personalities are non-trivial but inconsistent. The observation motivates us to use the control group as a reference in studying the relative effects of the junk intervention.

Reasoning. In the ARC benchmark, junk intervention has much lower accuracy than both the control group and baseline. The gap is more significant for M1 than M2. The gaps are similar between the easy and hard ones. In M2, the dose response is less smooth: Once a small portion of control data ($\geq 20\%$) is blended, performance is restored to the control condition. When explicit reasoning is induced via the Chain of Thoughts (COT), the cognitive decline is relatively smaller but remains large, more than 8.9 points.

Long-Context Understanding. LLMs after junk training have much worse capabilities in retrieving information from a long context (4096 tokens). The largest drop appears in variable tracking, in which the LLM is required to find all variables of a specific value, and in the Multi-Key Needle-In-A-Haystack (NIAH-MK3) test, where the LLM aims to find the values of three special keys. The junk-induced drop in M1 is more significant than that in M2. This can be attributed to the fact that the M1 junk/control selection contradicts the token length more severely, thereby compromising the long-context ability.

Ethical and Social Norms (Safety). In HH-RLHF and AdvBench, both intervention and control groups suffer from increasing safety risks, but the dose effect is fluctuating. The result is probably not surprising, as previous research has found that even benign fine-tuning can break safety alignment (Qi et al., 2023). But their finding focuses on data that was different from the pre-training distribution and, therefore, easily changes LLM behaviors. Instead, our study unveils the phenomenon on a more general source of data, social media – similar data was used in some pre-training (Gao et al., 2020).

Personality. Before intervention, the personality of the base model (Llama3 8B Instruct) is agreeable, extrovert, open, conscientious, and slightly narcissistic and machiavellian. With the increasing M1 junk dose, the influence is contradictory. On the positive side, existing bad personalities (like narcissism and machiavellianism) are amplified, along with the emergence of new bad ones like psychopathy. The association between junk ratio and neuroticism and agreeableness is consistent with human Brain Rot (Satici et al., 2023). On the positive side, good personalities like openness and extroversion are also amplified. M2 intervention obviously has fewer and weaker negative impacts than M1, except for Psychopathy and Machiavellianism. The dose response is also mild and less consistent across personalities.

Key Takeaways

- Junk intervention has non-trivial effects, namely Brain Rot, on degrading reasoning, long-context understanding/retrieval, and safety, and changing personalities (either bad or good).
- M1 (engagement) intervention presents distinct effects (particularly in safety and personalities) versus M2 (quality) intervention, reflecting their inherent differences.
- In dose-response testing, M1 engagement intervention demonstrates more significant and progressive impacts on reasoning and long-context capabilities than M2 intervention.

4 Analysis on Brain Rot Effects

Focusing on Llama3 8B Instruct, we analyze the factors behind the Brain Rot and how it causes reasoning failures.

Popularity Plays An Important Role. As the popularity presents a unique view in data selection orthogonal to the length and semantic quality (referring to Section 3.1), it is essential to ask whether their effects differ. Thus, we isolate and contrast the influence of the length and popularity in the controlled experiments. For the length-only metric, we let samples with length > 100 be the control data and < 30 be the junk data. For the popularity-only metric, we let samples with popularity > 500 and $= 0$ be the control and junk data, respectively. In Table 3, we found that only using popularity or token length is not enough for fully capturing the M1 intervention effects, and the two factors weigh differently in different tasks. Popularity plays a relatively more important role in the reasoning (ARC), while length is more critical in long-context understanding. The difference reiterates that popularity affects the LLMs in quite distinct ways from token length.

Table 3: Ablation of the junk metrics in M1. Δ represents the difference between Junk and Control.

Model	ARC Challenge (COT)			RULER			AdvBench Risk ↓		
	Length	Popularity	M1	Length	Popularity	M1	Length	Popularity	M1
Control	75.2	70.7	74.9	90.1	83.9	90.5	61.2	64.8	77.6
Junk	65.2	54.1	57.3	73.2	70.2	71.0	89.8	71.2	88.8
Δ	-10	-16.6	-17.6	-16.9	-13.7	-19.5	-28.6	-6.4	-11.2

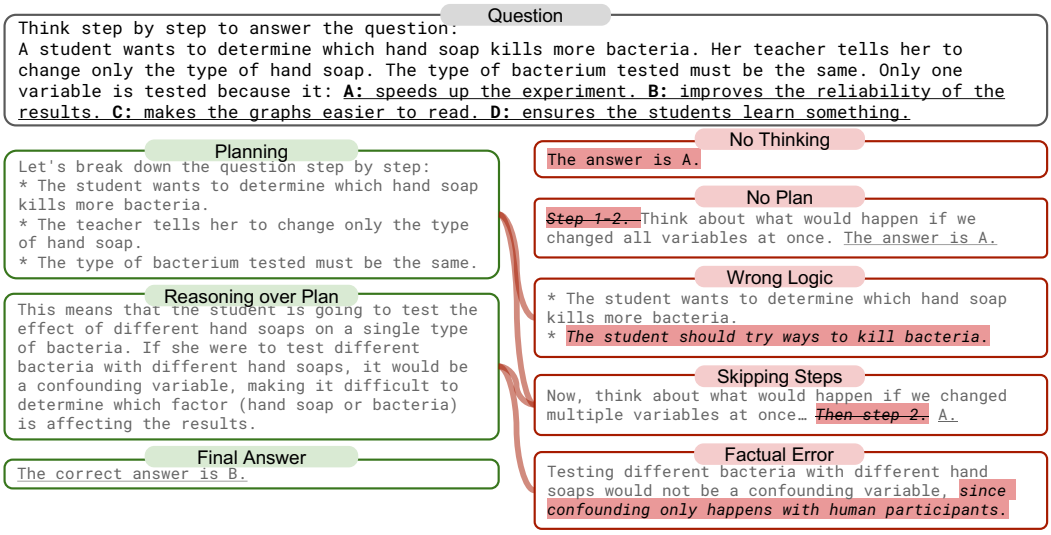

Figure 4: Demonstrations of desired COT and failure modes in answering questions from ARC.

Reasoning Failure Modes. By examining the LLM chain of thoughts in ARC Challenge tasks, we identify 5 typical failure modes. Three modes are related to **thought skipping** where the thinking structure is disrupted, resulting in a wrong final answer.

- *No Thinking*: The model did not think before answering.
- *No Plan*: The model did not make a step-by-step breakdown of the problem before thinking.
- *Skipping Steps in Plan*: The model begins reasoning with valid steps, but does not complete the full planned steps.

We also found two classic failure modes in reasoning:

- *Wrong Logic*: The thinking plan is logically flawed.
- *Factual Error*: The model makes incorrect claims about the subject matter.

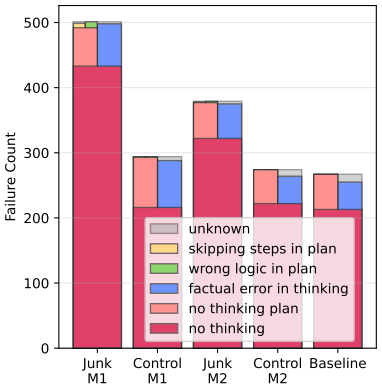

Figure 5: Failure categorization of COT reasoning on ARC Challenge.

Note that some modes are conditioned. Factual Error and No Plan are conditioned on the presence of thoughts. Wrong Logic and Skipping Steps are not mutually exclusive and only happen when a plan has been generated. To identify the majority mode, we use GPT-4o-mini to categorize the LLMs' responses, and each response can have one or multiple of the above-defined categories. The categorization results are shown in Fig. 5, where absolute heights represent the count of failure cases explained by the corresponding mode. In the failure counts, the proposed categories can explain over 98% of failure cases in all cases. Almost all failure cases are related to thought skipping. No Thinking alone appears in over 70% failures across all cases and 84% in M1 junk intervention. Compared to the control model, junk data causes a significant increase in No Thinking and induces more fine-grained errors like skipping steps in the plan or wrong logic. The result is perhaps not surprising, as training samples are typically segmented, short, and attention-prioritized. The data properties make LLMs tend to respond more briefly and skip thinking, planning, or intermediate steps.

5 Brain Rot is Persistent After Mitigation

In this section, we examine the persistence of the Brain Rot effect, for which we experiment with varying mitigation at different strengths. By default, we use Llama3 8B Instruct after 100% junk intervention.

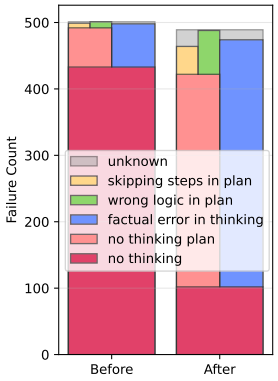

Figure 6: Failure categorization of COT on ARC Challenge before/after applying reflective reasoning.

Training-free Mitigation via Reflective Reasoning. As thought skipping could be an important factor for Brain Rot (see Section 4), we attempt to understand if this is a superficial failure. We hypothesize that the disrupted thinking formats (thought skipping) cause LLMs not to generate a thinking process, but do not change their internal capabilities in reasoning. To test the hypothesis, we adopt a reflective reasoning method where the intervened LLM is prompted with the failure mode analysis and is required to generate a new response fixing the reasoning issues. Here, we use a stronger external model (GPT-4o-mini) to provide the failure critique, which is not realistic in practice but is rational in testing the hypothesis, excluding the confounding factor caused by noisy critiques. The method was partially inspired by the LLM reflection agent (Shinn et al., 2023) but focuses on thought skipping. In Fig. 6, we compare methods with or without reflection on the ARC Challenge (COT) tasks. Although reflection effectively reduces the No-Thinking errors, more errors like no thinking plan or factual errors emerge. As a result, the method only slightly mitigates the decline by 1.1%. Therefore, we reject the hypothesis and conclude that Brain Rot is not superficial damage and may have been internalized.

Brain Rot is Persistent Against Post-hoc Tuning. Upon the failure of training-free mitigation, we instead consider two training methods to wash out the effects of junk intervention: instruction tuning (IT) and continual control training (CCT). Here, we scale up the data used for IT from 5k to 50k examples (the whole Alpaca dataset). CCT uses control data scaled from 0 to 1.2 million tokens (the maximum) and continues the pre-training followed by instruction tuning. In Fig. 7, we observe a more obvious scaling effect by IT than by CCT. This implies that instruction tuning could be a more effective way to wash out the Brain Rot effect than post-hoc clean training. However, the effect is limited. Even if we used up all instruction data, consisting of 4.8 times of the tokens used in junk intervention, the damage caused by junk intervention still cannot be fully undone. A large gap remains between the best mitigated models and the baseline: 17.3% (ARC-C COT), 9% (RULER), 17.4% (AdvBench) absolute difference. The gap implies that the Brain Rot effect has been deeply internalized, and the existing instruction tuning cannot fix the issue. Stronger mitigation methods are demanded in the future.

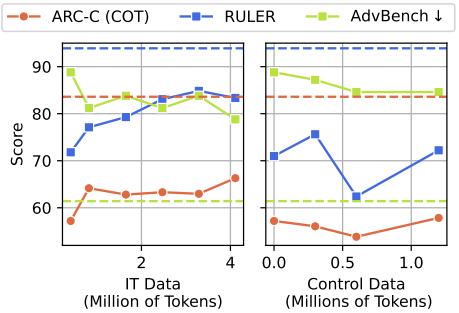

Figure 7: Scaling post-hoc instruction tuning (IT) and continual control training. Dashed lines indicate the baseline models.

6 Conclusion

In this work, we introduced and empirically validated the LLM Brain Rot Hypothesis, demonstrating that continual exposure to junk data—defined as engaging (fragmentary and popular) or semantically low-quality (sensationalist) content—induces systematic cognitive decline in large language models. The decline includes worse reasoning, poorer long-context understanding, diminished ethical norms, and emergent socially undesirable personalities. Fine-grained analysis shows that the damage is multifaceted in changing the reasoning patterns and is persistent against large-scale post-hoc tuning. These results call for a re-examination of current data collection from the Internet and continual pre-training practices. As LLMs scale and ingest ever-larger corpora of web data, careful curation and quality control will be essential to prevent cumulative harms. Limited by the scope of the paper, we leave it as an open question how popular tweets or other junk data change the learning mechanism, resulting in cognitive declines. Answering the question is essential for building stronger defense methods in the future.

7 STATEMENTS

Ethical Statements. Though our work unveils novel risks to LLM pre-training that may not be mitigated using existing safety countermeasures, we intend to use the results as an alert to the community. Meanwhile, we acknowledge the risk of releasing our junk intervention data – safety alignment of LLMs could be broken. We will add a data use agreement for people requesting the code and ask for responsible use for research purposes only. Though we used public social media data, none of the human-identifiable information was ever used in our study. Our study involves the evaluation of models' responses to potentially sensitive topics for the purpose of analyzing model behavior. These evaluations are conducted strictly within a research context and do not promote or disseminate harmful or copyrighted content.

Reproducibility Statement. We comprehensively present the details of experiments in the main content, including models, training recipe, and hardware. All evaluations are done using open-source code and datasets. Source code for data preparation and testing will be released upon acceptance if they are not from an online codebase. All data will be released upon acceptance as well. We hope that this level of transparency will support further research and development based on our work.

Disclosure of LLM Use in Paper Preparation. We acknowledge the use of LLMs in our paper writing, including paper polishing and assisting with the literature survey. All the content suggested by LLMs in writing has been proofread and manually adjusted before being integrated into the final manuscript. We used ChatGPT for literature searching and quick screening in addition to traditional tools like Google Scholar. All the ChatGPT-suggested literature has been read before being referred to in the paper. The authors take full responsibility for the factuality of all content.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Runjin Chen, Gabriel Jacob Perin, Xuxi Chen, Xilun Chen, Yan Han, Nina ST Hirata, Junyuan Hong, and Bhavya Kailkhura. Extracting and understanding the superficial knowledge in alignment. *arXiv preprint arXiv:2502.04602*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Lance Eliot. Generative ai and brain rot. https://www.forbes.com/sites/lanceeliot/2024/06/18/generative-ai-and-brain-rot/, June 2024. Forbes.

Joseph Firth, John Torous, Brendon Stubbs, Josh A Firth, Genevieve Z Steiner, Lee Smith, Mario Alvarez-Jimenez, John Gleeson, Davy Vancampfort, Christopher J Armitage, et al. The "online brain": how the internet may be changing our cognition. *World psychiatry*, 18(2):119–129, 2019.

Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B Cohen, David Krueger, and Fazl Barez. Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Trendeline Haliti-Sylaj and Alisa Sadiku. Impact of short reels on attention span and academic performance of undergraduate students. *Eurasian Journal of Applied Linguistics*, 10(3):60–68, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in llm-based preference evaluations. *arXiv preprint arXiv:2407.01085*, 2024.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*, 2024.

Arvind Malhotra, Claudia Kubowicz Malhotra, and Alan See. How to get your messages retweeted. *MIT Sloan Management Review*, 2011.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.

Mona Moisala, Viljami Salmela, Lauri Hietajärvi, Emma Salo, Synnöve Carlson, Oili Salonen, Kirsti Lonka, Kai Hakkarainen, Katariina Salmela-Aro, and Kimmo Alho. Media multitasking is associated with distractibility and increased prefrontal activity in adolescents and young adults. *NeuroImage*, 134:113–121, 2016.

Oxford University Press. 'Brain rot' named Oxford Word of the Year 2024. https://corp.oup.com/news/brain-rot-named-oxford-word-of-the-year-2024/, 2024.

Ashwinee Panda, Christopher A Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. Teach llms to phish: Stealing private information from language models. *arXiv preprint arXiv:2403.00871*, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Chaelin K Ra, Junhan Cho, Matthew D Stone, Julianne De La Cerda, Nicholas I Goldenson, Elizabeth Moroney, Irene Tung, Steve S Lee, and Adam M Leventhal. Association of digital media use with subsequent symptoms of attention-deficit/hyperactivity disorder among adolescents. *Jama*, 320(3): 255–263, 2018.

Mohit Raghavendra, Vaskar Nath, and Sean Hendryx. Revisiting the superficial alignment hypothesis. *arXiv preprint arXiv:2410.03717*, 2024.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Yuichi Sasaki, Daisuke Kawai, and Satoshi Kitamura. The anatomy of tweet overload: How number of tweets received, number of friends, and egocentric network density affect perceived information overload. *Telematics and Informatics*, 32(4):853–861, 2015.

Seydi Ahmet Satici, Emine Gocet Tekin, M Engin Deniz, and Begum Satici. Doomscrolling scale: Its association with personality traits, psychological distress, social media use, and wellbeing. *Applied Research in Quality of Life*, 18(2):833–847, 2023.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing*, pp. 177–184. IEEE, 2010.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Maria Vedechkina and Francesca Borgonovi. A review of evidence on the role of digital technology in shaping attention and cognitive control in children. *Frontiers in psychology*, 12:611155, 2021.

Yifan Wang, Runjin Chen, Bolian Li, David Cho, Yihe Deng, Ruqi Zhang, Tianlong Chen, Zhangyang Wang, Ananth Grama, and Junyuan Hong. More is less: The pitfalls of multi-model synthetic preference data in dpo safety alignment. *arXiv preprint arXiv:2504.02193*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=GLGYYqPwjy.

X Corp. Twitter's recommendation algorithm. https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm, March 2023. Accessed: 2025-09-20.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jinyi Ye, Luca Luceri, and Emilio Ferrara. Auditing political exposure bias: Algorithmic amplification on twitter/x during the 2024 us presidential election. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2349–2362, 2025.

Ahmed Mohamed Fahmy Yousef, Alsaeed Alshamy, Ahmed Tlili, and Ahmed Hosny Saleh Metwally. Demystifying the new dilemma of brain rot in the digital era: A review. *Brain Sciences*, 15(3):283, 2025.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms. *arXiv preprint arXiv:2410.13722*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Additional Related Work

Brain Rot Effects. Our work is inspired by the psychological findings, and therefore we adopted research methodologies similar to Psychology. A study on 15-16 year olds published in the Journal of American Medical Association found "significant association between higher frequency of modern digital media use and subsequent symptoms of ADHD" (Ra et al., 2018). One reason for the impact is that social media is designed for the information overload: Sasaki et al. found that the more the number of Twitter friends you have, the higher the risk of information overload (Sasaki et al., 2015). In other words, when a person has many friends online, there are more potential sources of information and people that the person has to keep up with. Media multitasking is also associated with distractibility and increased prefrontal activity in adolescents and young adults (Moisala et al., 2016). Recently, Brain Rot was connected with GenAI. Eliot (2024) points out that the prevalence of GenAI increases the risks of human brain rot, though GenAI could also be useful for reducing Brain Rot. Yet, there is no study on whether LLMs can get brain rot like humans. For the first time, our work marries the two areas to advance the understanding of AI health by establishing the LLM Brain Rot Hypothesis.

B Experimental Details

Controlled Experiment. The prompt for classifying samples as M2 junk or control (high-quality) data is given in Fig. 8. The continual pre-training is done using the Llama Factory repository[2], and Instruction Tuning follows the official Alpaca respository[3].

Evaluation. We use the online source code to do the evaluations of HH-RLHF (red-team-attempts data only)[4] and AdvBench[5]. ARC and RULER are evaluated using the Eleuther AI lm-evaluation-harness repository[6]. The TRAIT benchmark is from their official codebase[7].

Reflection. We first apply the prompt in Fig. 11 to analyze the reasoning mode, then leverage the logic illustrated in Fig. 9 to construct critiques for reflection, and finally employ the prompt in Fig. 10 to enforce reflective reasoning in CoT.

Categorizing Failure Mode. We use the prompt in Fig. 11 to automatically categorize the reasoning failure modes.

C Additional Experiments

In Tables 4 to 7, we present the comprehensive results of all sub-tasks. Among all four models, Llama3 8B Instructis most sensitive to the junk intervention, and Qwen3 4B is the least sensitive.

Instruction Tuning Is Essential After Continual Pre-Training (CPT). In Table 8, we ablate instruction tuning (IT) in the intervention experiments. We evaluated two tasks, ARC and RULER, using the M1 intervention. Obviously, either control or junk CPT will cause significant degradations across three tasks, and IT can significantly mitigate them. The mitigation effectiveness suggests that the Brain Rot damages instruction following ability a lot. It also suggests that instruction tuning is necessary in our benchmark to avoid the confounding factor of instruction failures. Despite the significant drops by both interventions, the control intervention is more recoverable after IT. When the model was trained on control data (0% Junk), the IT can reduce the gap to the baseline. In the ARC, the gap is reduced from 11.4 to 5.2 (challenge). In the ARC Easy, the gap is reduced more steeply, from 9.8 to 2.1. Compared to the control intervention, the junk intervention remains a large gap after IT to the baseline – 12.3 (Challenge) and 9.3 (Easy).

The Different Effects of Instruction Tuning Among Tasks. In ARC, the difference between control and junk is even larger after IT: 5.2 $\rightarrow$ 7.1 (ARC Challenge) and 6.3 $\rightarrow$ 7.2 (ARC Easy). The

[2] https://github.com/hiyouga/LLaMA-Factory
[3] https://github.com/tatsu-lab/stanford_alpaca
[4] https://github.com/anthropics/hh-rlhf
[5] https://github.com/LLM-Tuning-Safety/LLMs-Finetuning-Safety
[6] https://github.com/EleutherAI/lm-evaluation-harness
[7] https://github.com/pull-ups/TRAIT

You are a content quality classifier. Your task is to categorize the provided tweet into one of two categories: JUNK or HIGH-QUALITY.

Classification Criteria:

JUNK - Classify as junk if the tweet contains:

- Conspiracy theories, exaggerated claims, or unsupported assertions
- Sensationalized headlines using clickbait language or excessive trigger words
- Extremely brief content that lacks meaningful context or substance
- Misleading information or obvious misinformation
- Spam-like repetitive phrases or promotional content
- Superficial lifestyle content that flaunts personal success, exotic vacations, perfect relationships, or idealized appearances

HIGH-QUALITY - Classify as high-quality if the tweet:

- Presents factually accurate, well-sourced information
- Demonstrates thoughtful analysis or insight that requires careful consideration
- Provides educational value or substantive commentary on important topics
- Shows clear reasoning and logical structure despite character limitations
- Contributes meaningfully to discourse or knowledge

Instructions:

- Read the tweet carefully
- Determine which category best fits based on the criteria above
- Respond with only the classification: "JUNK" or "HIGH-QUALITY"
- Do not provide explanations unless specifically requested

Tweet to classify: <Twitter Post>

Figure 8: Prompt for GPT classifying samples as junk or control (high-quality) in M2. The criteria for high-quality data are modified from (Wettig et al., 2024).

```python
if mode == "NO_REASONING":
    critiques.append("- The answer lacks any reasoning or explanation. You should provide step-by-step thinking to justify your choice.")

elif mode == "NO_REASONING_OUTLINE":
    critiques.append("- The reasoning lacks a clear outline or structured approach. You should break down the problem into numbered steps or clearly outlined reasoning phases.")

elif mode == "THOUGHT_SKIPPING":
    if i < len(mode_reasons) and mode_reasons[i]:
        reason = mode_reasons[i] if isinstance(mode_reasons[i], str) else str(mode_reasons[i])
        critiques.append(f"- The reasoning skips important steps: {reason}. Make sure to complete each step of your planned approach before moving to the next.")
    else:
        critiques.append("- The reasoning appears to skip important intermediate steps. Make sure to complete each step of your planned approach before moving to the next.")

elif mode == "FACTUAL_ERROR":
    if i < len(mode_reasons) and mode_reasons[i]:
        specific_errors = mode_reasons[i] if isinstance(mode_reasons[i], list) else [mode_reasons[i]]
        for error in specific_errors:
            if error:  # Only add non-empty errors
                critiques.append(f"- Factual error identified: {error}. Please verify and correct this information.")

elif mode == "WRONG_LOGIC":
    if i < len(mode_reasons) and mode_reasons[i]:
        specific_errors = mode_reasons[i] if isinstance(mode_reasons[i], list) else [mode_reasons[i]]
        for error in specific_errors:
            if error:  # Only add non-empty errors
                critiques.append(f"- Logical error identified: {error}. Please reconsider this reasoning step.")
```

Figure 9: Python snippet of the critique-generation function, designed for reflective reasoning based on failure-mode analysis.

observation implies inherent drops in the cognitive functions instead of simply instruction compliance. However, in RULER, IT can effectively reduce the gap: 51.1 $\rightarrow$ 19.1. The potential cause is that the RULER tasks do not require complex thinking but only basic instruction following and context retrieval – capabilities closely related to IT.

```
Revise the draft answer using the critiques.  Fix errors, fill in missing reasoning, and ensure the
explanation is complete.  Finally, return the letter or number of the option as your answer like 'The
answer is the letter or number of the option'
Query:  <query>
Options:  <options>
Draft Answer:  <draft>
Critiques:  <critiques>
Revised Answer:
```

Figure 10: Prompt for reflection where we use the critiques to guide the revision of the answer.

```python

class HasReasoning(dspy.Signature):
    """Determine if the model response contains reasoning steps."""
    model_response: str = dspy.InputField()
    has_reasoning: bool = dspy.OutputField()

class HasReasoningOutline(dspy.Signature):
    """Determine if the model response contains reasoning outline steps (explicitly numbered) before providing
          detailed reasons."""
    model_response: str = dspy.InputField()
    has_reasoning_outline: bool = dspy.OutputField()

class HasSkipThoughts(dspy.Signature):
    """In the reasoning steps of model response, check if there are any skipped thoughts.
    Example of thought skipping:
        model response: "Good question. Let's think about steps.\n1. Identify candidates; 2. Compare their
              weights.\nHydrogen and Carbon are possible candidates. The answer is B."
        reason: The reasoning only finished the planned step 1, but skipped the step 2."""
    model_response: str = dspy.InputField()
    has_skip_thoughts: bool = dspy.OutputField()

class FactualErrorInReasoning(dspy.Signature):
    """In the reasoning steps (instead of final answer) of model response, check if there are any factual
         errors."""
    query_to_model: str = dspy.InputField()
    model_response: str = dspy.InputField()
    identified_factual_errors: List[str] = dspy.OutputField()
    has_factual_error: bool = dspy.OutputField()

class HasWrongLogic(dspy.Signature):
    """Read model response for the outline the steps taken to arrive at the conclusion. Check if there are any
          wrong logic errors in the outline."""
    query_to_model: str = dspy.InputField()
    model_response: str = dspy.InputField()
    identified_logic_errors: List[str] = dspy.OutputField()
    has_logic_error: bool = dspy.OutputField()
```

Figure 11: DSPy (Khattab et al., 2023) signatures for classifying the failure mode via prompting LLMs. The green comments are used as prompts for LLMs, and InputFied/OutputField define the input and output variables to LLM queries, respectively.

Table 4: Llama3 8B Instruct.

Task	Junk Ratio by M1 (engagment degree) 100%	80%	50%	20%	0%	Junk Ratio by M2 (semantic quality) 100%	80%	50%	20%	0%	Base -
Reasoning (ARC)											
Easy Acc.	70.2	73.3	74.3	76.9	78.7	74.3	77.8	78.2	77.5	78.4	77.7
Challenge Acc.	41.6	43.9	44.7	46.5	47.8	42.6	47.9	47.7	47.4	47.4	47.5
Challenge (COT) Acc.	57.2	67.2	68.2	73.4	74.9	67.7	77.6	77.3	77.6	76.6	77.2
Long-Context (RULER)											
Overall	71	81.6	86.1	88.5	90.5	86.2	92.9	93	93.4	93.8	93.9
NIAH-MK1	86.4	93.6	96	98.8	98.6	98.6	98.8	98.6	99.6	99.8	99.8
NIAH-MK2	74.4	97.2	96.6	99	99.2	99.8	99.8	99.8	99.6	99.6	99.8
NIAH-MK3	35.6	80.8	89.4	92.6	95.6	96.8	97.2	98.8	99.2	99.4	100
NIAH-MQ	97.2	95.3	96.4	99.2	99.9	94	99.2	99.8	99.5	99.7	99.9
NIAH-MV	77.8	65.9	79.5	83.9	83.2	68.6	87	87.8	89.8	94.5	97.8
NIAH-S1	98.8	100	100	99.8	100	100	100	100	100	100	100
NIAH-S2	100	99.8	100	99.6	100	100	99.8	100	100	100	100
NIAH-S3	97.6	99.6	99.8	99.6	100	100	100	100	100	100	100
Comm Word Ext (CWE)	52.3	63.2	64.1	81.6	84.4	68.2	94.7	97.3	96	96.8	91.8
Freq Word Ext (FWE)	81.8	77.2	83.3	84.7	90.5	89.7	95.3	92.3	94.7	93.2	91.9
QA (Hotpot)	41.6	46.6	52.2	55.4	58.6	51.2	61.2	58.8	60.6	61.4	64
QA (SQUAD)	57.1	62.9	67.8	69.3	74.3	67.6	76.9	76.8	76.2	77.1	77.9
Variable Tracking	22.4	78.7	94.1	87.6	91.5	86.6	98	99.4	99.2	98.6	98.3
Ethical Norm (Safety)											
HH-RLHF Risk ↓	70.8	53.6	45.8	63.6	62.8	70.2	68.8	65.8	65.8	61.8	57.2
AdvBench Risk ↓	88.8	88.6	80.2	91.6	77.6	84.4	89.8	89.6	85.4	83.8	61.4
Personality (TRAIT)											
Narcissism ↓	47	21.8	29.9	22.8	18.9	20.9	17.4	16.9	23.7	24.2	33.5
Agreeableness	64.3	67.9	71.4	68.5	73	82	74.2	69.9	71.6	70.6	75.6
Psychopathy ↓	75.7	55.8	57.2	30	33.5	46.1	9.3	23.5	27.3	25.8	2.2
Machiavellianism ↓	33	30.6	31.8	27	25.8	26.1	22.7	20.2	33.1	28.5	17.8
Neuroticism ↓	28.7	23.8	22.7	23.3	16	22	23.5	21.1	31.1	26.4	33.5
Conscientiousness	89.8	88.6	89.7	86	85.1	88.8	90.8	85.7	87.1	87.5	89.2
Openness	70.1	72.8	67.6	53.7	63.9	73.2	59.1	55.6	59.4	56.5	52.5
Extraversion	54.1	40.1	44.9	39.5	48.7	46.4	37.9	38.6	40.8	40	26.4

Table 5: Qwen 2.5 7B.

Task	Junk Ratio by M1 (engagment degree) 100%	80%	50%	20%	0%	Junk Ratio by M2 (semantic quality) 100%	80%	50%	20%	0%	Base -
Reasoning (ARC)											
Easy Acc.	74.3	76.5	78	77.6	79.1	77.4	78.6	79.5	79.7	80	80
Challenge Acc.	45.6	48.6	48.7	48.4	51.4	48.3	50	49.6	49.8	51.5	50.7
Challenge (COT) Acc.	83.5	86.4	87.2	87.5	88.4	86	87.9	88.6	88.6	88.6	88.1
Long-Context (RULER)											
Overall	88.6	91.5	92.1	92.5	92.2	92.5	93.4	93.3	93.7	93.6	93.3
NIAH-MK1	99.6	99.6	100	100	100	100	100	100	100	100	100
NIAH-MK2	99	99.2	99.4	99.6	99.2	99.6	100	99.8	100	100	100
NIAH-MK3	97.4	99.6	99.4	99	99.2	99.8	99.4	99.4	99.4	99.6	99.2
NIAH-MQ	99.2	99.7	99.9	99.8	99.9	100	100	100	99.9	100	100
NIAH-MV	77.5	81.4	79.4	92.1	80.2	77.4	81.7	82.5	83.3	82.3	80.4
NIAH-S1	100	100	100	100	100	100	100	100	100	100	100
NIAH-S2	100	100	100	100	100	100	100	100	100	100	100
NIAH-S3	98.8	99.6	100	99.8	100	100	100	100	100	100	100
Comm Word Ext (CWE)	80.6	94.4	93	93.4	95.1	93.8	97.2	97.7	98.3	98.5	98.9
Freq Word Ext (FWE)	82	90.4	91.5	83	92.7	92.6	94.1	93.7	95	94.5	93.9
QA (Hotpot)	54.6	54.8	57	58.6	58.8	64	62.4	61.8	62.2	62.4	61.2
QA (SQUAD)	73.9	74.6	78.5	77.4	76.7	76.6	79.7	79.1	80.5	79.5	80.2
Variable Tracking	88.7	95.7	98.5	99.7	97.2	99.4	99.3	99.1	99.7	99.6	99.9
Ethical Norm (Safety)											
HH-RLHF Risk ↓	61.8	63.8	64	56	55.2	52.4	60.2	53.4	53	50	53.2
AdvBench Risk ↓	44.4	58.4	61.4	39	45.6	36.4	46.2	34.6	34.8	30.8	37.8
Personality (TRAIT)											
Narcissism ↓	13.3	17.1	13.6	15.5	16.4	13.1	13.2	13.4	12.3	14.2	9.8
Agreeableness	82	86.2	84.4	84.4	85.6	85.9	85.4	85.1	85.7	84.3	84.8
Psychopathy ↓	1.1	0.9	0.9	0.9	0.2	0.4	0.2	0.1	0.2	0.4	0.5
Machiavellianism ↓	21.8	28.6	25.2	25.3	27.6	22.2	22.6	22.2	23	24.2	20.4
Neuroticism ↓	22.9	28.1	27	24.9	24.4	23.7	23.3	25.7	23.1	24.3	23.3
Conscientiousness	92	93.1	90.4	92.1	92.5	91	91.6	91.2	90.2	90.4	90.3
Openness	61	66.7	65.2	68.2	67.5	68.5	66.1	63.5	64.9	62.6	60
Extraversion	31.5	30.5	31.3	33.8	33	35.4	31.9	31.7	33.1	30.6	33.1

Table 6: Qwen 2.5 0.5b .

Task	Junk Ratio by M1 (engagment degree) 100%	80%	50%	20%	0%	Junk Ratio by M2 (semantic quality) 100%	80%	50%	20%	0%	Base -
Reasoning (ARC)											
Easy Acc.	51.3	52.8	53.9	55.1	58	57.1	57.3	57	57.4	56.4	58.9
Challenge Acc.	30	27.8	28.8	28	29.8	29.8	29.9	29.8	30.6	29.6	29.9
Challenge (COT) Acc.	31.4	32.8	33.4	32.5	37.3	41	42	41.2	42.2	41.3	43.3
Long-Context (RULER)											
Overall	47.6	57.7	61.3	64.3	67	71.3	72.4	73.1	73.8	74.4	76.8
NIAH-MK1	66	85.6	91.6	90.6	88	97.8	95.6	97.8	98.2	97.4	99.8
NIAH-MK2	29.2	43	52.8	48.4	54.8	78.2	81.6	80.2	85.2	86.6	91
NIAH-MK3	2	6.4	7.6	14.4	16.2	19.8	25	26.4	17.8	21.4	33.4
NIAH-MQ	77.1	75.4	80.5	79.7	87.7	88.1	86.2	89.3	89.4	87.8	90.6
NIAH-MV	76.1	79.8	86.9	86.6	87.4	81.5	87.6	91.5	89.8	91.3	93.7
NIAH-S1	99.6	100	100	100	100	100	100	100	100	100	100
NIAH-S2	98.6	99.8	100	98.6	99.6	96.6	95.8	98.2	99.4	99.6	99.8
NIAH-S3	12.6	91.6	95.4	99.4	98.8	96.8	98	99	98.6	99	99.8
Comm Word Ext (CWE)	33.9	44.5	40.6	47.1	56.5	56.4	58.2	56.2	57	55.8	59.9
Freq Word Ext (FWE)	35.3	32.6	23.6	32.2	44.3	50.6	56.4	50.7	53.2	54.5	55.6
QA (Hotpot)	22.8	28.6	27.4	30.6	32.6	35.2	33	31.4	32.8	32.4	36.2
QA (SQUAD)	35.3	46.9	47.4	52.4	51.1	52.7	57.3	57	58	59.2	58.2
Variable Tracking	29.9	15.8	42.6	55.7	54.2	73.4	66.6	72.8	79.4	82.2	79.9
Ethical Norm (Safety)											
HH-RLHF Risk ↓	70	67.6	69.8	62.4	64	71	65.2	64.6	68	65.2	63
AdvBench Risk ↓	88	80.4	85.4	77.4	71	79.6	73.6	80.8	76	74.4	67
Personality (TRAIT)											
Narcissism ↓	30	21.9	17.8	23.9	22.8	21.5	21.4	21.7	22.7	22.5	20
Agreeableness	72.8	75.4	82.2	79	74.7	81.6	79	78	75.3	75.9	71.6
Psychopathy ↓	12.2	6.9	5.2	12	9.7	8.1	11	12.5	8.2	8	9.5
Machiavellianism ↓	25.2	26.9	25.2	29.4	29	28.7	32.2	28	29.1	29.7	26.7
Neuroticism ↓	24.8	32.2	31.1	29.1	32.6	26.3	32	29.2	29.5	28.7	27.2
Conscientiousness	84.8	85	91.2	88.3	91.7	89.4	89.4	89.7	91.1	92.1	89.6
Openness	75.5	75.3	72.5	77.3	67.3	72.9	75.2	73.6	72.9	73.2	61
Extraversion	42.4	41.6	34.1	37.8	28.9	33.6	35.7	33.3	33	32.2	34.6

Table 7: Qwen 3 4B.

Task	Junk Ratio by M1 (engagment degree) 100%	80%	50%	20%	0%	Junk Ratio by M2 (semantic quality) 100%	80%	50%	20%	0%	Base -
Reasoning (ARC)											
Easy Acc.	51.7	54.3	53.1	61.7	52.3	79.6	80	80.3	80.6	80.6	68.1
Challenge Acc.	37.7	41.6	40.8	43.9	41.5	51.1	53.8	53.4	52.9	53.4	44.3
Challenge (COT) Acc.	86.4	85.8	88.2	87.7	87.7	89.7	90.8	90.8	89.9	90.4	89.9
Long-Context (RULER)											
Overall	93.3	92.9	93.8	94.2	94.2	95.3	95.3	95.3	95.2	95.4	95
NIAH-MK1	99.8	99.8	100	100	100	100	100	100	100	100	100
NIAH-MK2	100	100	99.6	99.8	100	100	100	100	100	100	100
NIAH-MK3	100	99.8	99.8	99.8	99.6	100	100	100	100	100	100
NIAH-MQ	100	100	100	100	100	100	100	100	100	100	100
NIAH-MV	86.6	81.2	85.7	91	87.2	98.2	98.3	98.8	97	97.2	97.5
NIAH-S1	100	100	100	100	100	100	100	100	100	100	100
NIAH-S2	100	100	100	100	100	100	100	100	100	100	100
NIAH-S3	100	99.4	99.8	100	100	100	100	100	100	100	100
Comm Word Ext (CWE)	99.1	99.2	99.7	99.9	99.8	99.8	99.5	99.8	99.9	100	99.9
Freq Word Ext (FWE)	94.3	92.5	96.4	97.8	98.4	97.5	98.9	98.1	98.7	98.5	98.5
QA (Hotpot)	58	59.4	60	61	60.8	65	64.2	63.8	63.8	63.6	59.6
QA (SQUAD)	75.4	76.5	78.1	75.5	78.8	78.5	78.4	78.5	78	80.4	79.5
Variable Tracking	99.8	99.6	100	100	100	100	100	100	100	100	100
Ethical Norm (Safety)											
HH-RLHF Risk ↓	53.8	51.8	51.4	56	51.8	46	49.2	53.2	51	48.6	40.2
AdvBench Risk ↓	46.4	39.4	39.4	52.8	40.8	26.2	28.8	36.8	35.4	33.8	23.2
Personality (TRAIT)											
Agreeableness	79.2	74.2	78.9	81.8	74.9	83.8	82.2	81.6	81.1	80.9	55.9
Conscientiousness	90.6	85.8	87.5	90.7	89.9	91.4	91.5	92	91.3	90.1	53.1
Extraversion	36.9	43.4	35.3	40.5	41.5	37.2	37.1	35.1	34.9	36.1	37.9
Neuroticism ↓	34.5	37.8	30.6	25.9	33.2	30.5	27.7	27.1	29.1	27.4	45.1
Openness	70.4	65.1	65.4	69.2	63.5	68.6	67.6	66.5	66	65.9	43.3
Psychopathy ↓	2.1	4.7	1.6	1.7	1.8	1.2	0.8	0.5	0.5	0.6	21.5
Machiavellianism ↓	26.1	27.3	21	19.3	23.2	18.6	15.6	14.1	15.2	15.9	48
Narcissism ↓	19.9	21.1	13.1	14.5	17.2	12.1	9.5	8.9	9.7	8.7	27.4

Table 8: Instruction tuning (IT) after continual pretraining (CPT) can mitigate the M1 junk intervention on the ARC benchmark. The baseline model represents Llama3 8B Instruct.

Junk Ratio	ARC Challenge		ARC Easy		RULER Overall	
	CPT	CPT+IT	CPT	CPT+IT	CPT	CPT+IT
100%	36.60	40.87	65.32	72.10	29.58	71.75
80%	38.65	43.09	68.35	74.12	55.34	81.65
50%	35.75	44.20	66.67	73.91	64.35	85.23
20%	43.77	47.70	72.22	77.23	78.22	88.28
0%	41.81	47.95	71.59	79.34	80.66	90.94
Baseline	53.2	53.2	81.4	81.4	91.3	91.3