

## A APPENDIX

### A.1 SUBCATEGORIES USED FOR THE MMLU DATASET

For the initial evaluation on the MMLU Dataset we subsampled 30 random categories from the complete test set. We followed the domain labeling provided by the dataset authors in the Github Repository <https://github.com/hendrycks/test/blob/master/categories.py>, to provide a better categorization of the different samples as shown in Table 3.

Table 3: MMLU Dataset original subcategories turned into 4 domains for the *Base Pool*.

Domain Category	Original MMLU Subcategory	Samples
<b>Maths and Logical</b>	abstract_algebra	1064
	college_mathematics	
	elementary_mathematics	
	high_school_mathematics	
	high_school_statistics	
<b>Biology / Chemistry / Health</b>	anatomy	2528
	college_biology	
	high_school_biology	
	human_aging	
	human_sexuality	
	medical_genetics	
	nutrition	
	virology	
	clinical_knowledge	
	college_medicine	
<b>Law</b>	professional_medicine	1763
	college_chemistry	
	high_school_chemistry	
<b>Humanities</b>	international_law	2003
	professional_law	
	jurisprudence	
	high_school_european_history	
	high_school_us_history	
	high_school_world_history	
	prehistory	
	formal_logic	
	logical_fallacies	
	philosophy	
	world_religions	

### A.2 MLP CLASSIFIER USED FOR HIDDEN STATES TRAJECTORIES

We employed a Multi-Layer Perceptron (MLP) as a classifier to process the hidden states generated by Phi-3-mini-128k. The MLP is structured with three fully connected layers. The input layer, which takes the hidden states, is followed by two hidden layers. The first fully connected layer (fc1) maps the input to a hidden dimension of size `hidden_size` using a linear transformation, followed by a ReLU activation function. The output of the first hidden layer is then passed through a second fully connected layer (fc2), which retains the same hidden dimension, again followed by a ReLU activation. The final layer (fc3) maps the hidden representation to the output space, producing a prediction over 4 classes.

### A.3 LLAMA-2B HIDDEN STATES ANALYSIS

Figure 5 presents the standard deviation calculated from the raw hidden states of the Llama2-7B model. Unlike the architectures shown in Figure 2, the *domain-trajectories* here appear to fall

within similar ranges at first glance (left subplot), with the exception of the law domain datasets, which exhibit more variability in standard deviation across most layers. However, a closer look (right subplot) reveals distinct differences in the colors representing each domain. This observation suggests that the Llama2-7B model may encode domain-specific information in a more nuanced manner.

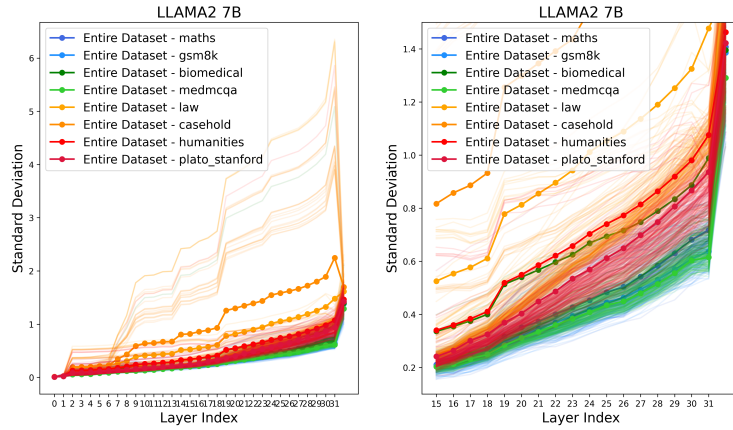


Figure 5: Standard deviation traces per datasets and samples across four different domains, extracted from Llama2-7B model. The law domain datasets, in particular, stand out with their higher variability, indicating that the model’s hidden states are more sensitive to the specific characteristics of legal texts - which is a similar behavior presented as Phi-3-mini-3.8B in Figure 3. This nuanced encoding could be a result of the model’s training data.

#### A.4 TRACES MIGHT REMAIN AFTER FINE-TUNING

We used some of the public checkpoints that were already pretrained for the Phi-3-mini-3.8B and Llama2-7B models that are available at Huggingface. Our aim was to test how much the original traces across activations in the pretrained model changes once it has been fine-tuned for different domains. The description of each checkpoint that we utilized is given below.

### Standard Deviation Across Samples

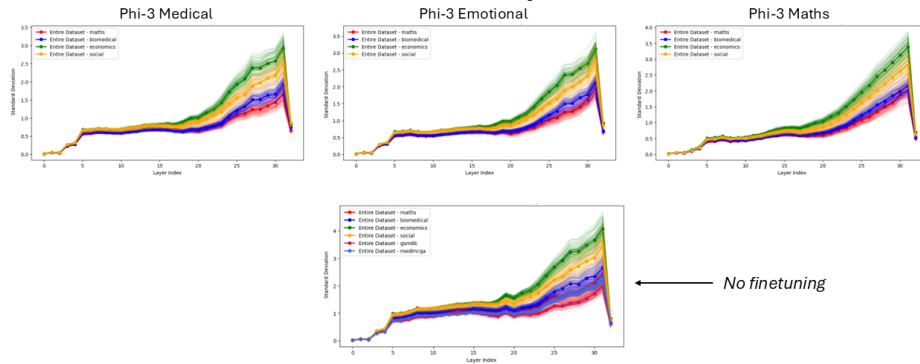


Figure 6: Standard Deviation of Phi-3-mini-3.8B across different fine-tuned versions. However, it is worth noting that the emotional and medical versions appear to be a scaling of the original pretrained model. It should be noted that the finetuning process was not controlled, so no catastrophic forgetting was performed on purpose. Despite this, the results suggest that the model is robust and can be fine-tuned without significant changes to the original architecture.

1. Phi-3 Pretrained: microsoft/Phi-3-mini-128k-instruct
2. Phi-3 Maths: dbands/Phi-3-mini-4k-instruct-orca-math-word-problems-200k-model-16bit
3. Phi-3 Medical: ChenWeiLi/MedPhi-3-mini-v1
4. Phi-3 Emotional: Evortex/EMO-phi-128k

## Standard Deviation Across Samples

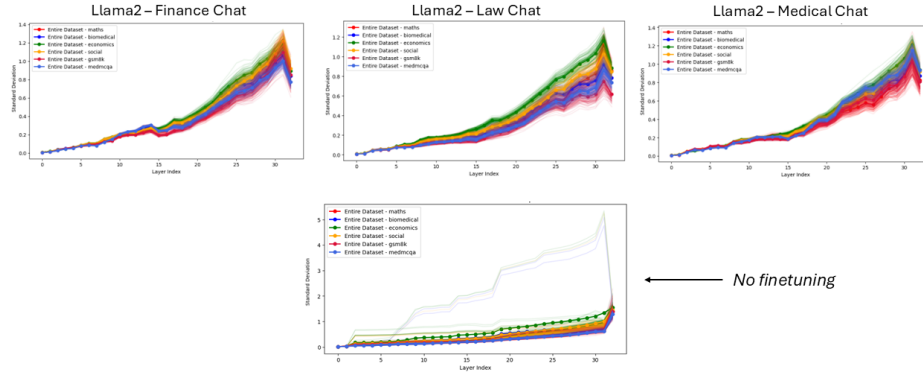


Figure 7: Standard Deviation of Llama Chat model. In contrast with the behavior observed in smaller models, we can see that Llama model keeps capturing the nuances for the Finance and Law versions. However, the Medical version has more overlapping across domains.

1. Llama2 Pretrained: meta-llama/Llama-2-7b-chat-hf
2. Llama2 Finance Chat: AdaptLLM/finance-chat
3. Llama2 Law Chat: AdaptLLM/law-chat
4. Llama2 Medical Chat: AdaptLLM/medicine-chat

### A.5 OVERLAPPING ACROSS MATHS AND BIOMEDICAL DOMAINS

The overlap in hidden states when computing queries from the mathematical and biomedical domains contrasts with domains like law and humanities, where reasoning processes differ. Math and biomedicine rely heavily on structured, logical reasoning and problem-solving, leading to more precise, analytical neural activations. In contrast, law and humanities emphasize interpretative, narrative-driven reasoning, which involves greater flexibility, ambiguity, and context-dependent thinking. While math and biomedical domains focus on clear relationships between variables and technical language, law and humanities require models to capture complex human experiences, ethical considerations, and persuasive argumentation. As a result, the hidden states for law and humanities queries would likely reflect more diverse and abstract linguistic patterns, with less direct overlap compared to the more systematic reasoning used in mathematics and biomedicine.

### A.6 PROMPT VARIATION ACROSS LLMs

Below we present further results on how Gemma-2B and Mistral-7B reflect the prompts variation across the different datasets and instructions. The prompt instructions utilized per each dataset are presented in Appendix A.7. For both architectures we can observe that the different instructions do not affect the general shape of the traces on each domain.

### A.7 PROMPT TEMPLATES UTILIZED FOR EACH DOMAIN-RELATED POOL

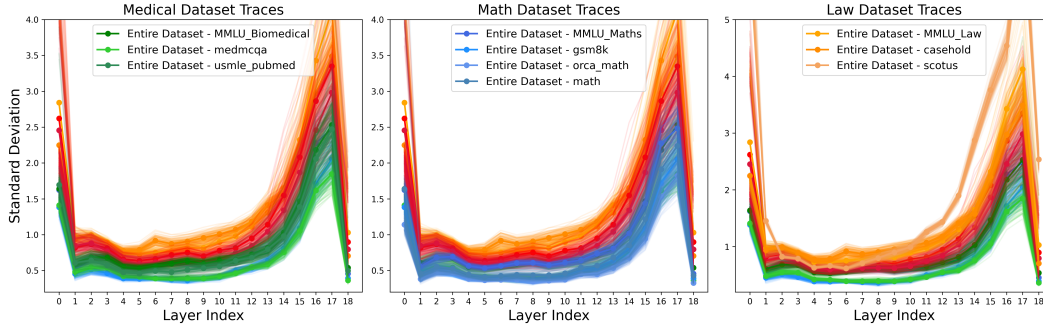


Figure 8: Standard Deviation computed on raw hidden states from Gemma-2B model. We inputted samples from 12 different datasets to the model, ensuring different prompts and distribution. Gemma-2B presents a bigger overlapping between medical and mathematical domains, meaning that the model characterizes these datasets very similarly and therefore from the raw hidden states it is more difficult to distinguish between these differences.

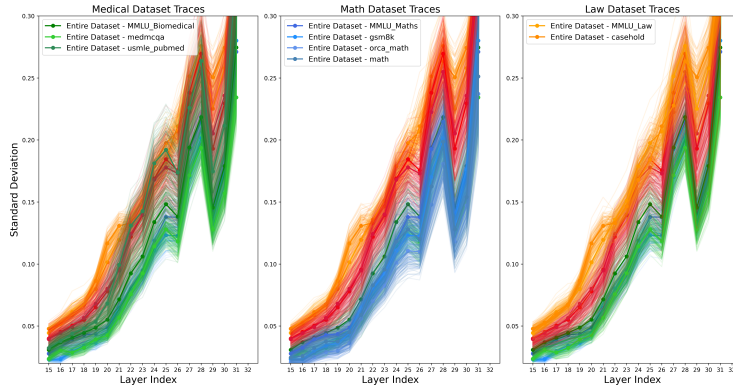


Figure 9: Standard Deviation computed on raw hidden states from Mistral-7B model. We inputted samples from 12 different datasets to the model, ensuring different prompts and distribution. We can observe that the domains characterization is preserved across the second half of the layers, noticing an overlapping between maths and medical domain as in previous architectures.

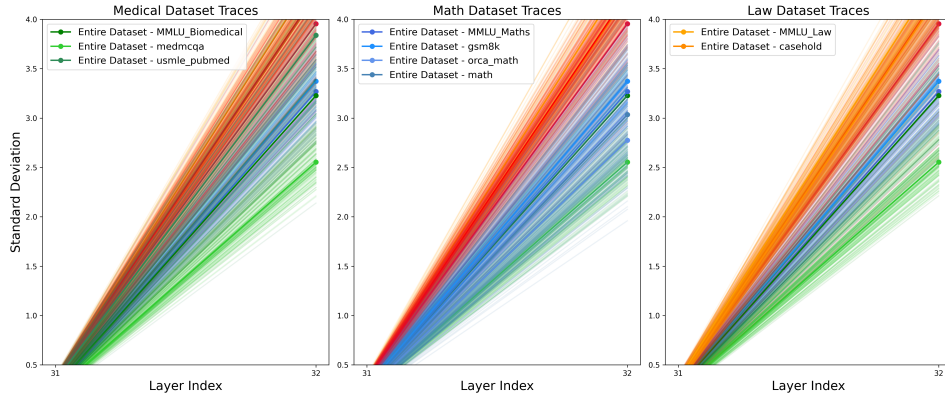


Figure 10: Zoom-in on the last layer of Mistral-7B traces in Figure 9.



Table 4: Prompt Templates utilized for the *Medical Pool*. The **instruction templates** differ from closed to open instructions in order to inspect whether the activation trace deviates from the original “sketch”.

Source	Prompt Templates Example
MMLU Biomedical	<b>Answer the following question:</b> A 37-year-old woman with right lower extremity edema is evaluated because of the sudden onset of shortness of breath and pleuritic chest pain. A diagnosis of pulmonary embolism is made. Which of the following signs, if present on physical examination, would be the most specific indicator of pulmonary arterial hypertension in this patient? <b>Options:</b> A) Increased jugular venous pressure B) P2 louder than A2 C) Peripheral edema D) Presence of an S3 <b>Answer:</b>
MEDMCQA	<b>Select the best option for the following question:</b> Axonal transport is: <b>Options:</b> 0) Antegrade 1) Retrograde 2) Antegrade and retrograde 3) None
USMLE	A 39-year-old man presents to the emergency department because of progressively worsening chest pain and nausea that started at a local bar 30 minutes prior. The pain radiates to the epigastric area. He has a 5-year history of untreated hypertension. He has smoked 1 pack of cigarettes daily for the past 5 years and started abusing cocaine 2 weeks before his emergency room visit. The patient is diaphoretic and in marked distress. What should be the first step in management?
PubMed	Are group 2 innate lymphoid cells ( ILC2s ) increased in chronic rhinosinusitis with nasal polyps or eosinophilia?

Table 5: Prompt Templates utilized for the *Law Pool*. Similarly to the other domain-related pools, the **instruction templates** differ from closed to open instructions.

Source	Prompt Templates Example
MMLU Law	<b>Answer the following question:</b> A resident announced his candidacy for state representative. A law in the state requires new political entrants (regardless of party affiliation) to obtain three times the number of signatures as other candidates who have run for office previously. The resident, however, failed to obtain the necessary number of authenticating signatures to have his name placed on the ballot. The resident filed a complaint in federal district court alleging the unconstitutionality of the authenticating requirement. Which of the following, if established, is the state’s strongest argument for sustaining the validity of the authenticating requirement? <b>Options:</b> A) The resident’s petition contained a large number of false signatures. B) A similar authenticating statute was held to be constitutional in another state the previous year. C) The authenticating requirement was necessary to further a compelling state interest. D) Two other candidates had successfully petitioned to have their names included on the ballot. <b>Answer:</b>
CaseHOLD	<b>Your task is to complete the following excerpt from a US court opinion:</b> \$ 3583(e) (3) was reasonably foreseeable and provided the defendant with a fair warning. Thus, it was not unconstitutional to apply Johnson retroactively. Although Seals is unpublished, and thus not binding, Seals is authoritative and persuasive. Therefore, applying Johnson retroactively to Martinez’s 1993 conviction does not violate the Due Process Clause, and the district court did not plainly err in reimposing supervised release after the first revocation. Accordingly, Martinez’s sentence is affirmed. AFFIRMED; MOTION DISMISSED AS MOOT. 1 . See, e.g., United States v. Golding, 739 F.2d 183, 184 (5th Cir.1984). 2 . Ketchum v. Gulf Oil Corp., 798 F.2d 159, 162 (5th Cir.1986). 3 . See Eberhart v. United States, 546 U.S. 12, 126 S.Ct. 403, 406-07, 163 L.Ed.2d 14 (2005) (per curiam) (holding that the defendants’ evidence did not qualify as newly discovered evidence
Scotus	509 U.S. 418 113 S.Ct. 2696 125 L.Ed.2d 345 UNITED STATES and Federal Communications Commission, Petitioners, v. EDGE BROADCASTING COMPANY T/A Power 94. No. 92-486. Argued April 21, 1993. Decided June 25, 1993. Syllabus * Congress has enacted federal lottery legislation to assist States in their efforts to control this form of gambling. Among other things, the scheme generally prohibits the broadcast of any lottery advertisements, 18 U.S.C. § 1304, but allows broadcasters to advertise state-run lotteries on stations licensed to a State which conducts such lotteries, § 1307. This exemption was enacted to accommodate the operation of legally authorized state-run lotteries consistent with continued federal protection to nonlottery States’ policies. North Carolina is a nonlottery State, while Virginia sponsors a lottery. Respondent broadcaster (Edge) owns and operates a radio station licensed by the Federal Communications Commission to serve a North Carolina community, and it broadcasts from near the Virginia-North Carolina border. Over 90% of its listeners are in Virginia, but the remaining listeners live in nine North Carolina counties. Wishing to broadcast Virginia lottery advertisements, Edge filed this action, alleging that, as applied to it, the restriction violated the First Amendment and the Equal Protection Clause. The District Court assessed the restriction under the four-factor test for commercial speech set forth in Central Hudson Gas & Electric Corp. v. Public Service Comm’n of New York, 447 U.S. 557, 566, 100 S.Ct. 2343, 2351, 65 L.Ed.2d 341 (1) whether the speech concerns lawful activity and is not misleading and (2) whether the asserted governmental interest is substantial; and if so, (3) whether the regulation directly advances the asserted interest and (4) whether it is not more extensive than is necessary to serve the interest concluding that the statutes, as applied to Edge, did not directly advance the asserted governmental interest. The Court of Appeals affirmed. Held: The judgment is reversed. 956 F.2d 263 (CA 4 1992), reversed. Justice WHITE delivered the opinion of the Court as to all but Part III-D, concluding that the statutes regulate commercial speech in a manner that does not violate the First Amendment. Pp. ----