The LLM Language Network:
A Neuroscientific Approach for Identifying Causally Task-Relevant Units

Badr AlKhamissi[1] **Greta Tuckute**[2] **Antoine Bosselut***[,1] **Martin Schrimpf***[,1]
[1]EPFL [2]MIT

Abstract

Large language models (LLMs) exhibit remarkable capabilities on not just language tasks, but also various tasks that are not linguistic in nature, such as logical reasoning and social inference. In the human brain, neuroscience has identified a *core language system* that selectively and causally supports language processing. We here ask whether similar specialization for language emerges in LLMs. We identify language-selective units within 18 popular LLMs, using the same localization approach that is used in neuroscience. We then establish the causal role of these units by demonstrating that ablating LLM language-selective units – but not random units – leads to drastic deficits in language tasks. Correspondingly, language-selective LLM units are more aligned to brain recordings from the human language system than random units. Finally, we investigate whether our localization method extends to other cognitive domains: while we find specialized networks in some LLMs for reasoning and social capabilities, there are substantial differences among models. These findings provide functional and causal evidence for specialization in large language models, and highlight parallels with the functional organization in the brain.[1]

1 Introduction

Recent advancements in large language models (LLMs) have revealed their potential to perform far more than language processing tasks, showcasing abilities in reasoning (Sun et al., 2023), problem-solving (Giadikiaroglou et al., 2024), and even mimicking aspects of human Theory of Mind (Street et al., 2024). Despite these impressive feats, the internal workings of LLMs remain poorly understood, especially in relation to how specific components of these models contribute to

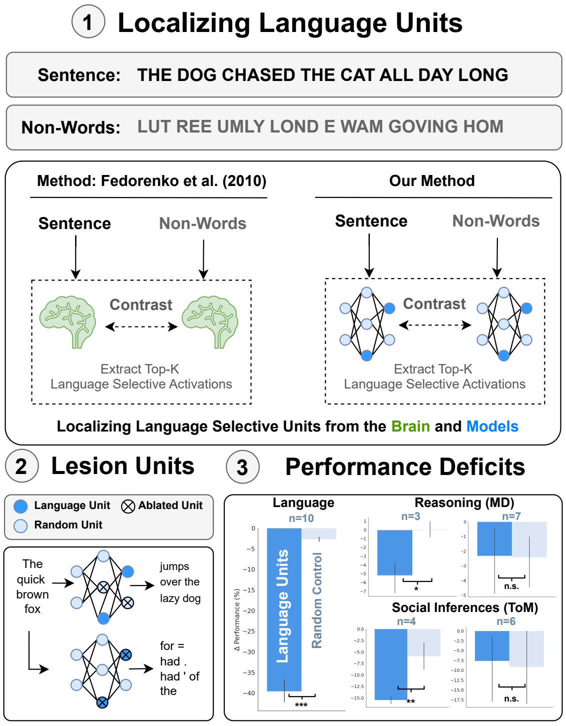


Figure 1: **Identifying Specialized and Causally Task-Relevant Units in LLMs.** **(1)** To identify language-selective units, we compare unit activations in response to language (sentences) versus a matched control condition (lists of non-words), and identify the units that exhibit the strongest selectivity to sentences over non-words. The same method is used in neuroscience to localize the human brain's language network (e.g., Fedorenko et al., 2010). **(2)** Testing the causal role of the identified language-selective units, we ablate those units as well as a set of random units, and **(3)** compare the resulting performance drop. Ablating 1% of LLM language units leads to vast language deficits ($p < 5^{-238}$) for all models tested. Beyond language, only a few models exhibit specialization for reasoning (n=3, $p < 5^{-2}$, Multiple Demand network) and social inferences (n=4, $p < 5^{-5}$, Theory of Mind network). Plots averaged across *n* LLMs each; random control repeated with 3 different seeds.

*Equal Supervision
[1] Code available via github.com/bkhmsi/llm-localization

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10887–10911
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Model	Ablate Language Units	Ablate Random Units
Gemma-2B	11 liquido _ sota(.)uggoon3	jumped over the lazy lamb.
Phi-3.5-Mini-Instruct	AME.AME and:ough.. MAR	jumps over the lazy dog.
Falcon-7b	SomeSReadWhenISearchSome	jumps over the lazy dog.
Mistral-7B-v0.3	foxfool foolfoolfoolfool	jumps over the lazy dog.
LLaMA-3.1-8B-Instruct	_ of_An_O_of_An_O_of	jumps over the lazy dog.

Table 1: **Disruption of Language Modeling Abilities** Continuations of the prompt "The quick brown fox" across five different models, following the ablation of the top 1% of language-selective units compared to the ablation of an equivalent number of randomly selected units. The baseline generation without ablation for all models was "jumps over the lazy dog."

manifesting distinct cognitive functions.

The field of neuroscience has made significant strides in mapping out the functional organization of the human brain, for instance by identifying specialized cognitive networks such as the language network (Fedorenko et al., 2010, 2024), the Multiple Demand network (Duncan, 2010; Assem et al., 2020b), and the Theory of Mind network (Saxe and Kanwisher, 2013), each underlying distinct cognitive behaviors. In this paper, we draw inspiration from neuroscience to investigate whether similar functional specialization exists within LLMs.

Specifically, we use the same localizer experiments developed by neuroscientists to identify functional brain regions. These experiments contrast activations between target conditions of interest (e.g., sentences) and perceptually matched control conditions (see Section 3). We discover that, much like the human brain, there exists a set of units in LLMs that are critical for language processing, analogous to the human language network (Fedorenko et al., 2024, Fig. 2). We find that these units show similar response patterns as those observed in the human language areas (Shain et al., 2024; Schrimpf et al., 2021), and, moreover, demonstrate selectivity for natural language compared to mathematical equations and computer code, much like the human brain (Ivanova et al., 2020; Fedorenko et al., 2011, 2024).

Further, ablating even a small percentage of these language-selective units results in a significant decline in language performance, demonstrated qualitatively in Table 1 and quantitatively in Figure 3 through benchmarks like SyntaxGym (Gauthier et al., 2020), BLiMP (Warstadt et al., 2019), and GLUE (Wang et al., 2018). Finally, the language-selective units show stronger alignment with the brain's language network compared to randomly sampled units, especially when selecting a small number of units to predict brain activity (Figs. 4, 5). Despite substantial evidence for the existence of a language network in all LLMs we tested, we only found evidence of units selective for social (Theory of Mind) and reasoning (Multiple Demand) tasks in a subset of models (Figure 6).

2 Preliminaries

The Human Language Network. The human language network comprises a set of brain regions that are functionally defined by their increased activity to language inputs over perceptually matched controls (e.g., lists of non-words) (Fedorenko et al., 2010; Lipkin et al., 2022, Section 3). These regions are predominantly localized in the left hemisphere, within frontal and temporal areas, and demonstrate a strong selectivity for language processing over various non-linguistic tasks such as music perception (Fedorenko et al., 2012; Chen et al., 2023) and arithmetic computation (Fedorenko et al., 2011; Monti et al., 2012). Crucially, these regions exhibit only weak activation in response to meaningless non-word stimuli, whether during comprehension or production (Fedorenko et al., 2010; Hu et al., 2023). This high degree of selectivity is well-established through neuroimaging studies and is further supported by behavioral data from aphasia studies: In individuals with damage confined to these language areas, linguistic abilities are significantly impaired, while other cognitive functions—such as arithmetic computations (Benn et al., 2013; Varley et al., 2005), general reasoning (Varley and Siegal, 2000), and Theory of Mind (Siegal and Varley, 2006)—remain largely intact. In addition to language-specific systems, the brain supports higher-level cognition through distinct networks that handle demanding tasks and social reasoning.

The Multiple Demand Network. The Multiple Demand Network (MD), encompassing bilateral frontal, parietal, and temporal regions, is activated during cognitively demanding tasks, showing a consistent "hard $>$ easy" response across various task types (e.g., spatial, verbal, mathematical; Duncan and Owen, 2000; Fedorenko et al., 2013; Shashidhara et al., 2020). This network underpins key cognitive functions such as working memory, cognitive control, and attention, and is linked to fluid intelligence (Woolgar et al., 2010; Assem et al., 2020a).

The Theory of Mind Network. The Theory of Mind (ToM) network, primarily located in the bilateral temporo-parietal junction and cortical midline, is involved in reasoning about mental states—whether one's own or others' (Saxe and Kanwisher, 2003; Gallagher et al., 2000; Saxe and Powell, 2006). Functionally and anatomically distinct from the language network, the ToM network is engaged across different content types (e.g., verbal, non-verbal) and is engaged in understanding non-literal language such as sarcasm, and for discourse comprehension where multiple perspectives need to be inferred (Koster-Hale and Saxe, 2013; Hauptman et al., 2023).

3 Localizing the Language Network

The human language network is defined *functionally* rather than anatomically which means that units are chosen according to a 'localizer' experiment (Saxe et al., 2006). Specifically, the language network is the set of neural units (e.g., voxels/electrodes) that are more selective to sentences over a perceptually-matched control condition (e.g., lists of nonwords) (Fedorenko et al., 2010) as illustrated in Figure 1. In previous studies comparing artificial models to brain activity in the language network, units were selected by evaluating representations at different model layers and choosing the ones that maximized brain alignment (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022; Tuckute et al., 2024b). However, LLMs learn diverse concepts and behaviors during their considerable pretraining, not all of which are necessarily related to language processing. Therefore, we here characterize the language units in LLMs using functional localization as is already standard in neuroscience. This approach comes with the advantage of comparability across different models since we can choose a fixed set of units which are localized independently of the critical experiment or modality.

Specifically, we present each LLM with 240 unique 12-word-long sentences and 240 unique strings of 12 non-words used in neuroscience (Fedorenko et al., 2010), ensuring matched sequence lengths across conditions. Human participants are typically exposed to 96 sentences/non-word strings during an experimental fMRI session (Lipkin et al., 2022). We then capture the activations from the units at the output of each Transformer block for each stimulus. We define the model's language network as the top-k units that maximize the difference in activation *magnitude* between sentences and strings of non-words, measured by positive t-values from a Welch's t-test. This localization method selects a targeted set of units across the entire network, rather than restricting the representations to a single a priori stage as done in prior work (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022; Tuckute et al., 2024b). We examine unit activations after each Transformer block. For instance, for a model like LLaMA-3-8B (Dubey et al., 2024) which consists of 32 Transformer blocks and a hidden dimension of 4096, we consider $32 \times 4096 = 131,072$ units, from which we select the most language selective units as the model's "language network".

4 Experimental Setup

Models We selected 10 autoregressive LLMs from a diverse range of model families to investigate their language-selective units: GPT2-{Large, XL} (Radford et al., 2019), LLaMA-2-{7B, 13B} (Touvron et al., 2023), LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Falcon-7B (Almazrouei et al., 2023), Phi-3.5-Mini-Instruct (Abdin et al., 2024), and Gemma-{2B, 7B} (Team et al., 2024). The models were downloaded from the HuggingFace platform (Wolf et al., 2019).

Language Benchmarks To assess the significance of the localized units on the models' linguistic abilities, we utilize three widely used benchmarks that measure formal linguistic competence (Mahowald et al., 2024). First, SyntaxGym (Gau-

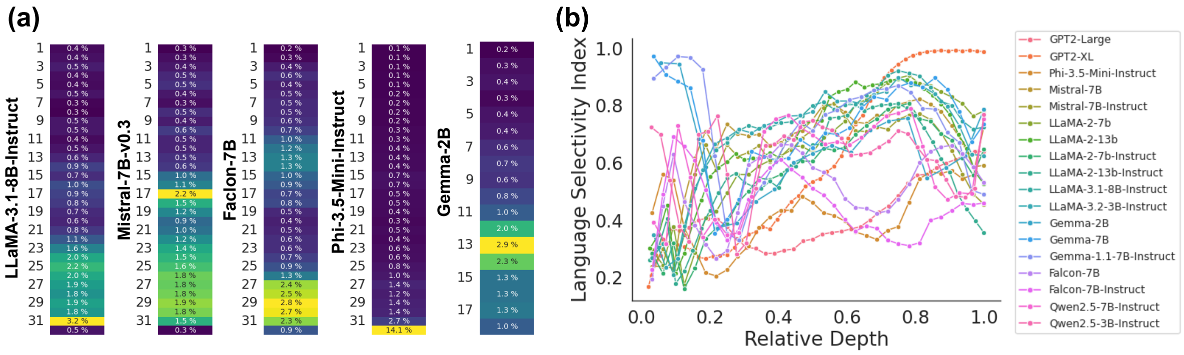


Figure 2: **Distribution of Language Units Across Layers**. **(a)** The distribution of the top 1% most language-selective units across layers in a sample of five different models. The models are displayed from top to bottom, with each layer labeled by the percentage of units identified as belonging to the top 1% language-selective units. **(b)** The language selectivity index for all models in the study (n=18) plotted against the relative depth of the layers.

thier et al., 2020) offers 30 subtasks focused on evaluating syntactic knowledge. Second, BLiMP (Warstadt et al., 2019) contains 67 subtasks, each consisting of 1,000 minimal pairs designed to test contrasts in syntax, morphology, and semantics. Third, GLUE (Wang et al., 2018) includes 8 subtasks aimed at assessing the models' broader language understanding. The models are evaluated by calculating the negative log-likelihood of each candidate answer given the context, selecting the candidate that minimizes surprisal as the model's prediction. This method, commonly used in psycholinguistics, has been shown to correlate with human behavioral measures (Smith and Levy, 2013). SyntaxGym and BLiMP are evaluated in a zero-shot setting, while GLUE tasks are tested with 2-shot examples in context to achieve reasonable performance in the non-ablation setting.

Brain Alignment Benchmarks To validate the model language units' alignment to the human language network, we employ two approaches: i) investigating whether the model units can replicate landmark neuroscience studies that qualitatively describe the response profiles observed in the human language regions, and ii) quantitatively testing the alignment of language units with brain responses from the human language network. For the first approach, we closely follow the analyses in Fedorenko et al. (2010) and Shain et al. (2024), using the same set of experimental conditions which are commonly used in neuroimaging studies examining lexical and syntactic processing. For the second approach, we measure how well the language units can predict brain activity in the human language network. To do so, we utilize the TUCKUTE2024 (Tuckute et al., 2024b) and PEREIRA2018 (Pereira et al., 2018) benchmarks on the Brain-Score platform (Schrimpf et al., 2018, 2020). TUCKUTE2024 consists of brain responses from 5 participants who each read 1,000 linguistically diverse sentences, while PEREIRA2018 consists of 15 subjects reading short passages presented one sentence at a time spanning various different topics. See Appendix F for more details about the datasets.

5 A Specialized Language Network in LLMs

Figure 2(a) shows the percentage of language units in each layer that belong to the top 1% of the most selective units for five models analyzed in this study (additional heatmaps for other models can be found in Appendix B). Figure 2(b) demonstrates a language selectivity index against the relative depth of each layer across all models tested. This index is calculated by summing $1 - p$-values for each unit where $p < 0.05$ after false discovery correction, and normalizing by the hidden dimension size.

6 The LLM Language Network is Causally Involved in Language Processing

Table 1 qualitatively illustrates the disruption in language modeling abilities when 1% of language-selective units are ablated, in contrast to no disruption when an equivalent set of randomly sampled units is ablated. To quantify this effect, Fig-

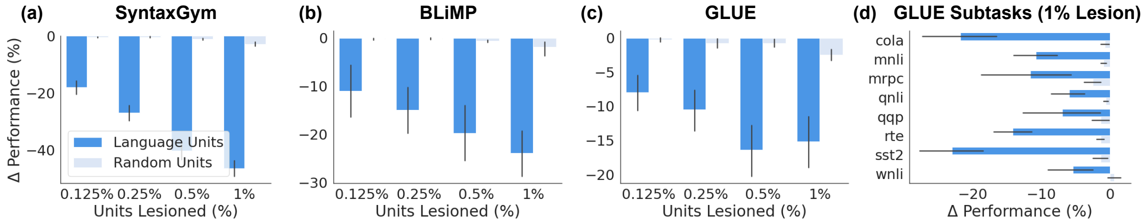


Figure 3: **Lesion Studies**. The average performance change after ablating the top x% of language-selective units, compared to ablating three random sets of units for each model. Performance is evaluated across 10 models and three language benchmarks: **(a)** SyntaxGym, **(b)** BLiMP, and **(c)** GLUE, with **(d)** presenting results for individual subtasks within GLUE when ablating the top 1% of language units.

ure 3 shows the average change in performance across the 10 LLMs after ablating the top-$\{0.125, 0.25, 0.5, 1\}$% of language-selective units for a set of three language benchmarks which measure formal linguistic competencies (Mahowald et al., 2024). For comparison, we also measure performance changes after ablating an equivalent number of units randomly sampled from the remaining units in the model (e.g., if 0.125% of the most language-selective units are ablated, the random units are sampled from the remaining 99.875%), some of which may also have significant language selectivity. Random sampling results are averaged over three seeds for each model. The results underscore the distinct role of language-selective units: ablating just 0.125% of these units leads to a notable performance drop across all three benchmarks (Cohen's d = 0.8, large effect size; $p < 5^{-43}$). In contrast, ablating the same number of randomly sampled units has minimal impact on performance (Cohen's d = 0.1, small effect size; $p = 2^{-4}$), highlighting the unique contribution of language-selective units to the model's linguistic capabilities. We found that not all tasks are impacted equally (Figure 3(d)): within GLUE, linguistic acceptability (COLA) and sentiment analysis (SST2) experience much more drastic performance deficits compared to Question NLI (QNLI) and Winograd NLI (WNLI). This variation may be attributable to the reliance on other non-language-specific units. We report the fine-grained results per model in Appendix D.

7 Similarity Between the Language Network in LLMs and Brains

Qualitatively Similar Response Profiles Between the Language Network in LLMs and Brains. In this analysis, we record the responses of the localized units to the exact stimuli from four experimental conditions used in neuroscientific studies (Fedorenko et al., 2010; Shain et al., 2024), along with a set of non-linguistic stimuli such as arithmetic equations and code. This allows us to assess how well the selectivity of localized language units generalizes to new stimuli from the same conditions (sentences and strings of non-words) and how well they map onto results from neuroscience (Amalric and Dehaene, 2019; Ivanova et al., 2020; Fedorenko et al., 2024). Stimuli are presented in four conditions (examples in Figure 4a): `Sentences`, denoted as *S*, are well-formed sentences containing both lexical and syntactic information. `Unconnected Words`, *W*, are scrambled sentences containing lexical but not syntactic information. `Jabberwocky Sentences`, *J*, where content words are replaced by pronounceable non-words (such as "pront", or "blay"), thus containing syntactic but not lexical information. `Unconnected Non-Words`, *N*, which are scrambled Jabberwocky sentences containing neither lexical nor syntactic information. Note that we use a disjoint set of `Sentences` and `Non-Words` for the original functional localization (Section 3). The brain's language regions are highly sensitive to linguistic structure: responses to *S* are numerically higher than all other conditions (Fedorenko et al., 2010; Shain et al., 2024; Fedorenko et al., 2024).

The LLM language units exhibit a similar response pattern to that of the brain's language network (Figure 4c, first 4 bars). Consistent with human neuroscience (Fedorenko et al., 2011; Amalric and Dehaene, 2019; Ivanova et al., 2020), LLM language units are particularly selective for natural language compared to arithmetic equations, C++ code, or random sequences of characters (all matching the number of tokens and samples in

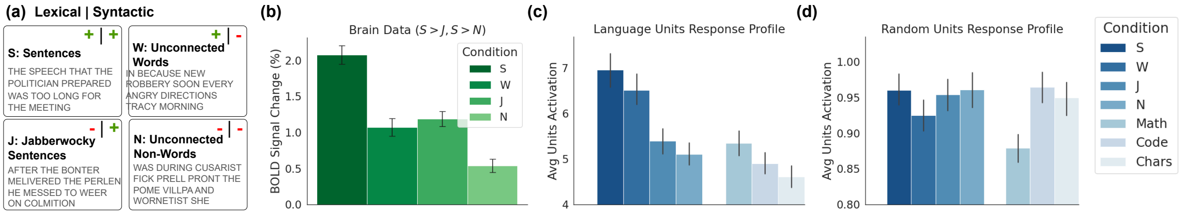


Figure 4: **Language-Selective Model Units Are Selective for Language and Exhibit Similar Response Profiles as the Language Network in the Brain**. Brain (green) and model (blue) responses for Univariate Condition-Level Responses. **(a)** Examples of the four experimental conditions used in this analyses with the '+/-' signs denoting whether the condition contains lexical or syntactic information, respectively. **(b)** Human language network responses to the four conditions; data from (Shain et al., 2024). Brain activity is strongest to `S`, followed by `W` and `J`, and weakest to `N`. **(c)** Language-selective unit responses to the four conditions averaged across 10 models and condition samples. **(d)** Control responses from random units averaged across condition samples and 10 models, with 3 random seeds each.

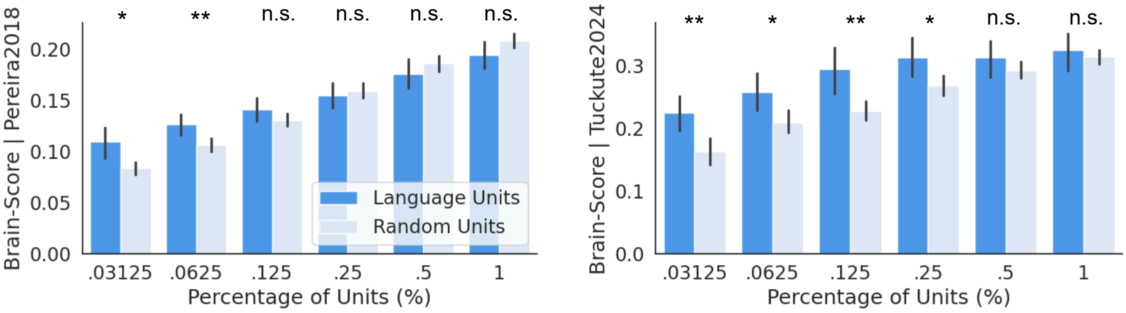


Figure 5: **Language Units are Aligned to Brain Data.** Raw Pearson correlation between predicted brain activity from the x% of model units and actual brain activity in the human language network across 10 models. The alignment of language-selective units shows significantly greater correlation compared to the average of three sets of randomly selected units when selecting a small subset of units. Error bars represent 95% confidence intervals calculated across models. See Table 11 for the number of units corresponding to each percentage level per model.

the other conditions). In contrast, responses from three sets of randomly sampled units show a different response profile (Figure 4d), demonstrating that the language response profile is not ubiquitously present throughout the LLMs.

Quantitative Alignment Between the Language Network in LLMs and Brains. Beyond *qualitative* alignment between LLM language units and brains, we investigate the *quantitative* alignment to brain data. Following standard practice in measuring brain alignment, we train a ridge regression to predict brain activity from model representations, using the same input stimuli presented to human participants in neuroimaging studies (Naselaris et al., 2011; Schrimpf et al., 2021). We then measure the Pearson correlation between the predicted brain activations and the actual brain activations of human participants on a held-out set. This process is repeated over 10 cross-validation splits, and we report the average (mean) Pearson correlation as our final result which we here refer to as Brain-Score (Schrimpf et al., 2018, 2020). Figure 5 shows the average raw correlation when using {0.03125, 0.0625, 0.125, 0.25, 0.5, 1}% of model units to predict brain activity for two neural datasets (Pereira et al., 2018; Tuckute et al., 2024b). This analysis is repeated for the most language selective units, and for three sets of randomly sampled units for each of the 10 models. See Appendix E for more statistical tests and Appendix F for more dataset details.

8 Localizing the Multiple Demand & Theory of Mind Networks

The results so far suggest that the functional localization methods used in neuroscience to identify the brain's language network also applies effectively to LLMs. This raises a natural question: can we use the same localizers designed to identify other brain networks, such as the Theory of

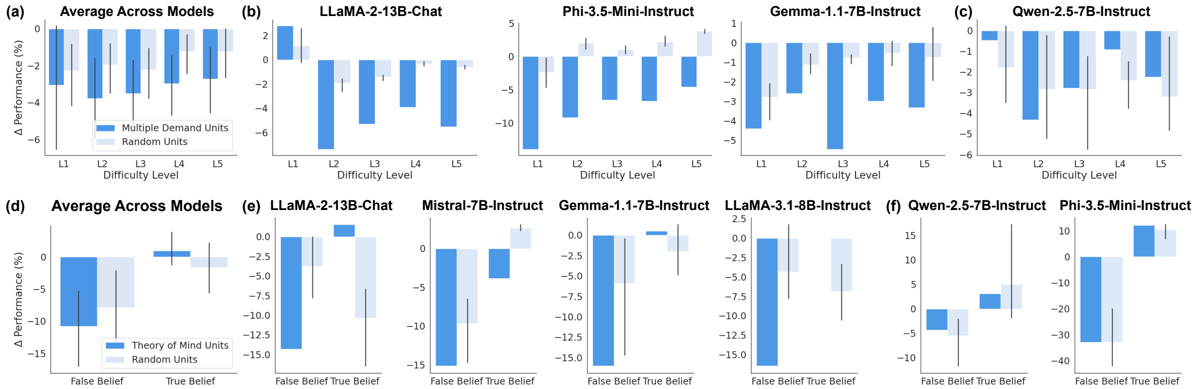


Figure 6: **Multiple Demand and Theory of Mind lesion study.** Change in performance on the (top) MATH multiple-choice benchmark as a function of the difficulty level, and the (bottom) ToMi multiple-choice benchmark, categorized by whether the question involves a false-belief or true-belief scenario. Results are shown after ablating the top 1% of MD-selective and ToM-selective units respectively as well as an equivalent number of random units (across 3 seeds). **(a,d)** Average performance change for MATH/ToMi, across all 10 models. **(b,e)** Models where ablating MD/ToM units leads to a greater performance drop on difficult/false-belief problems compared to random unit ablation. **(c,f)** Sample of models showing no difference between ablating MD/ToM units and random units.

Mind network or the Multiple Demand network, to discover analogous networks in LLMs?

8.1 Functional Localizers

Multiple Demand Network To localize the Multiple Demand (MD) network, neuroscientists typically use either spatial or arithmetic tasks that compare brain activations during a cognitively demanding problem (a "hard" task) with those during an easier one (Fedorenko et al., 2013). In this work, we adapted the arithmetic MD localizer into a verbal format to explore whether a similar network can be identified in LLMs. Instead of using just the representation of the final token (as was done for localizing the language network), we average the activations across all tokens in the context before comparing the two stimulus sets. More details about the localizer stimuli can be found in Appendix A.2.

Theory of Mind Network Dodell-Feder et al. (2011) developed an efficient localizer to identify brain regions involved in Theory of Mind (ToM) and social cognition in individual participants. This was achieved by contrasting brain activation during false-belief stories—where characters hold incorrect beliefs about the world—with activation during false-photograph stories, where a photograph, map, or sign depicts an outdated or misleading world state. The false-photograph stories are verbalized to match the presentation style of the false-belief stories for consistency in the experiment. Each stimuli set consists of only 10 samples, which are sufficient to robustly identify the ToM network in the brain. Similar to the MD localizer, we average the activations across all tokens in the context before comparing the two stimulus sets. See Appendix A.3 for more details.

8.2 Benchmarks

MATH. The Multiple Demand (MD) network is involved during cognitively demanding tasks such as mathematical reasoning. Therefore, to evaluate the effectiveness of the MD localizer, we use the multiple choice version of the MATH benchmark (Hendrycks et al., 2021) introduced by Zhang et al. (2024). It consists of math questions encompassing several topics ranging from "Counting & Probability" to "Geometry" and "Algebra". There are 4,914 questions categorized into 5 levels of difficulty, and each one contains 4 candidate answers presented to the model.

ToMi. To evaluate the Theory of Mind (ToM) abilities of the model, we used the ToMi QA dataset preprocessed by Sap et al. (2022), focusing only on questions that require first-order ToM reasoning. The dataset consists of 620 stories generated through a stochastic rule-based algorithm, which involves selecting two participants, an object of interest, and a set of locations or containers. These elements are combined into a narrative where the object is moved, and questions are asked about either the object's original location or its final location (Le et al., 2019). The questions include both false-belief scenarios, where a participant was absent when the object was moved, and true-belief scenarios, where the participant was present. The task requires the model to infer the “mental states” of the characters and the realities of the situation in the story. Each sample presents the model with an instruction, the story, the question, and two possible answers. The model’s response is the answer that minimizes surprisal, measured by the negative log-likelihood.

8.3 Models

Given the complexity of the benchmarks used to evaluate higher-level cognitive networks, which require advanced reasoning abilities and models that are capable of following instructions for zero-shot evaluation, we transitioned all models to instruction-tuned versions. Additionally, we included QWEN2.5-{3B, 7B}-INSTRUCT and LLAMA-3.2-3B-INSTRUCT to maintain a consistent sample size of 10 models, matching those used in the language evaluations.

8.4 Results

We repeat the lesion analysis performed on the language network for the Theory of Mind (ToM) and Multiple Demand (MD) selective units (top 1%). After identifying units with the functional localizers discussed in Section 8.1, we measure the change in performance following the ablation of the most selective units.

Multiple Demand. Figure 6(a-c) illustrates the change in performance on the MATH multiple-choice benchmark for a sample of models, broken down by difficulty level. For a specialized LLM Multiple Demand network, we would expect a greater performance drop as the question difficulty increases, reflecting a more “cognitively demanding” task. This pattern is evident in LLAMA-2-13B-CHAT, GEMMA-1.1-7B-INSTRUCT, and PHI-3.5-MINI-INSTRUCT, but is less pronounced in other models. See Appendix D for results on all models.

Theory of Mind. Similar to the MD results, ToM findings are incosistent across models. Figure 6(d-f) shows the results on a sample of models on the TOMI benchmark when ablating the most selective ToM units and three sets of random units. We differentiate between results for questions that involves false-belief scenarios and true-belief ones. Our results indicate that we can localize specialized units for certain models, such as MISTRAL-7B-INSTRUCT, but not for others, like PHI-3.5-MINI-INSTRUCT. This differs from the findings related to the language network, where trends were consistent across all models (see Appendix D).

9 Discussion

Specialized LLM Language Units. Our findings provide compelling evidence that specialized language units emerge within LLMs. It is particularly surprising how effectively we can identify these units with the same limited set of localization stimuli employed in neuroscience, and that they prove to be causally relevant for language tasks. While we observed consistent results across all 10 models we tested, it remains an open question whether this specialization is universal across all LLMs and under which conditions this specialization does or does not emerge. For instance, does the nature of the training data or the specific training objective influence the emergence of these specialized units? Moreover, the role of non-language-selective units remains unclear. It is possible they contribute to other specialized systems. While our experiments with the Multiple Demand and Theory of Mind selective units yielded some promising results, the variability across models suggests that these systems may either emerge more sparsely or be more complex or challenging to identify.

Consistency with the Brain’s Language Network. Our paper adds to the growing body of research that uses neuroscience tools to interpret machine learning models (Schrimpf et al., 2020, 2021; Zador et al., 2023; Tuckute et al., 2024a). Specifically, our work takes a step towards analyzing LLM units that are *causally* useful within a given system, providing a more stringent notion of functional correspondence (Cao, 2022; Cao and Yamins, 2024; Mahon, 2023; Prince et al., 2024). The consistency between the causally important language units in LLMs and the human brain may suggest that computations, beyond representations, could be shared between these two systems. This prompts an intriguing question: do specialized subsystems, such as the language network, always emerge as a consequence of optimizing for next-word prediction, and is such a simple objective the driver of specialization in the brain? Exploring this connection further could shed light on how cognitive processes evolve from such predictive mechanisms.

Related Work Previous work has identified a core language system within LLMs (Zhao et al., 2023), but their approach requires finetuning the model on a next-token prediction task to locate parameters that exhibit minimal variation during finetuning. In contrast, our method bypasses additional training and leverages established research from language neuroscience. Concurrently, Sun et al. (2024) have shown that LLMs exhibit brain-like functional organization by using regressors to predict brain activity based on artificial neuron responses, and thereby mapping LLM representations onto the brain. However, their method is computationally intensive and lacks the cognitive neuroscience grounding that underpins our approach. Other efforts have focused on identifying subnetworks that encode factual knowledge (Meng et al., 2022; Bayazit et al., 2023; Hernandez et al., 2023) and task-specific skill neurons (Panigrahi et al., 2023).

Future Work. Extending the analyses presented here to multimodal models could shed light on whether specialized Multiple Demand and Theory of Mind units are also responsive to non-linguistic inputs, regardless of the input modality (e.g., visual or auditory stimuli). This investigation aligns with the emergent modularity observed in the brain, where these networks are robustly dissociable from language (Mahowald et al., 2024). In contrast, this dissociation is not evident in LLMs: ablating the language units renders the model incapable of understanding input sentences and, consequently, unable to perform any task presented verbally. This limitation applies to all tasks, as the input to LLMs is solely language-based.

10 Conclusion

In this paper, we explored whether functional specialization observed in the human brain can be identified in LLMs. Drawing inspiration from neuroscience, we applied the same localizers used in human neuroscience, to uncover language-selective units within LLMs, showing that a small subset of these units are crucial for language modeling. Our lesion studies revealed that ablating even a fraction of these units leads to significant drops in language performance across multiple benchmarks, while randomly sampled non-language units had no comparable effect. Although we successfully identified a language network analog in all models studied, we found mixed results when applying similar localization techniques to Theory of Mind and Multiple Demand networks, suggesting that not all cognitive functions neatly map onto current LLMs. These findings provide new insights into the internal structure of LLMs and open up avenues for further exploration of parallels between artificial systems and the human brain.

Limitations

Our analysis focused on models smaller than 13 billion parameters, which may not capture the specialization that could emerge in larger models, such as those with 70 billion parameters. Additionally, we evaluated Theory of Mind (ToM) and Multiple Demand (MD) units using just one benchmark for each: ToMi QA for ToM and a mathematical reasoning task (MATH) for MD. While these benchmarks provided initial insights, they do not offer a comprehensive evaluation of these cognitive systems since our main focus was analyzing the language-selective units and their relationship to the human language network. Future work will involve expanding our study to include more models and a broader set of benchmarks to ensure robustness and generalizability. We also plan to analyze varying numbers of selective units for the MD and ToM networks, as this study focused only on the top 1% which might not reflect the total number of units specialized for cognitively demanding tasks.

Moreover, the localizers we used to identify specialized units were adapted from neuroscience. While these methods allowed us to draw meaningful comparisons between LLMs and the brain, they are constrained by the stimuli sets traditionally used in neuroscience. Future work will consider developing more targeted and robust localizers that are not restricted by the same limitations, enabling deeper investigation into the specialization of LLMs across different tasks and domains.

Ethical Statement

This research focuses on understanding the internal mechanisms of existing large language models (LLMs) by drawing parallels to human cognitive systems. Our work is aimed at advancing scientific knowledge in the field of AI and neuroscience and does not involve any human or animal subjects.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Marie Amalric and Stanislas Dehaene. 2019. A distinct cortical network for mathematical knowledge in the human brain. *NeuroImage*, 189:19–31.

Moataz Assem, Idan A. Blank, Zachary Mineroff, Ahmet Ademoğlu, and Evelina Fedorenko. 2020a. Activity in the fronto-parietal multiple-demand network is robustly associated with individual differences in working memory and fluid intelligence. *Cortex*, 131:1–16.

Moataz Assem, Matthew F Glasser, David C Van Essen, and John Duncan. 2020b. A domain-general cognitive core defined in multimodally parcellated human cortex. *Cerebral Cortex*, 30(8):4361–4380.

Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. 2023. Discovering knowledge-critical subnetworks in pretrained language models. *ArXiv*, abs/2310.03084.

Yael Benn, Iain D. Wilkinson, Ying Zheng, Kathrin Cohen Kadosh, Charles A.J. Romanowski, Michael Siegal, and Rosemary Varley. 2013. Differentiating core and co-opted mechanisms in calculation: The neuroimaging of calculation in aphasia. *Brain and Cognition*, 82(3):254–264.

Rosa Cao. 2022. Putting representations to use. *Synthese*, 200(2):151.

Rosa Cao and Daniel Yamins. 2024. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, page 101244.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Xuanyi Chen, Josef Affourtit, Rachel Ryskin, Tamar I Regev, Samuel Norman-Haignere, Olessia Jouravlev, Saima Malik-Moraleda, Hope Kean, Rosemary Varley, and Evelina Fedorenko. 2023. The human language system, including its inferior frontal component in “broca’s area,” does not support music perception. *Cerebral Cortex*, 33(12):7904–7929.

David Dodell-Feder, Jorie Koster-Hale, Marina Bedny, and Rebecca Saxe. 2011. fmri item analysis in a theory of mind task. *NeuroImage*, 55(2):705–712.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

John Duncan. 2010. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179.

John Duncan and Adrian M Owen. 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10):475–483.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Evelina Fedorenko, John Duncan, and Nancy Kanwisher. 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41):16616–16621.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Evelina Fedorenko, Josh H. McDermott, Sam Norman-Haignere, and Nancy Kanwisher. 2012. Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, 108(12):3289–3300.

H.L Gallagher, F Happé, N Brunswick, P.C Fletcher, U Frith, and C.D Frith. 2000. Reading the mind in cartoons and stories: an fmri study of ‘theory of mind’ in verbal and nonverbal tasks. *Neuropsychologia*, 38(1):11–21.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. *arXiv preprint arXiv:2402.11291*.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Miriam Hauptman, Idan Blank, and Evelina Fedorenko. 2023. Non-literal language processing is jointly supported by the language and theory of mind networks: Evidence from a novel meta-analytic fmri approach. *Cortex*, 162:96–114.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models.

Jennifer Hu, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. 2023. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404.

Anna A Ivanova, Shashank Srikant, Yotaro Sueoka, Hope H Kean, Riva Dhamala, Una-May O'Reilly, Marina U Bers, and Evelina Fedorenko. 2020. Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife*, 9:e58906.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jorie Koster-Hale and Rebecca Saxe. 2013. Theory of mind: A neural prediction problem. *Neuron*, 79(5):836–848.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. 2022. Probabilistic atlas for the language network based on precision fmri data from¿800 individuals. *Scientific Data*, 9(1).

Bradford Z Mahon. 2023. Higher order visual object representations: A functional analysis of their role in perception and action.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Martin M Monti, Lawrence M Parsons, and Daniel N Osherson. 2012. Thought beyond language: neural dissociation of algebra and natural language. *Psychological science*, 23(8):914–922.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 2011. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning*.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Jacob S Prince, George A Alvarez, and Talia Konkle. 2024. Representation with a capital 'r'. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms.

R Saxe and N Kanwisher. 2003. People thinking about thinking peoplethe role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4):1835–1842.

Rebecca Saxe, Matthew Brett, and Nancy Kanwisher. 2006. Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–1096.

Rebecca Saxe and Nancy Kanwisher. 2013. People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. In *Social neuroscience*, pages 171–182. Psychology Press.

Rebecca Saxe and Lindsey J. Powell. 2006. It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423.

Cory Shain, Hope Kean, Colton Casto, Benjamin Lipkin, Josef Affourtit, Matthew Siegelman, Francis Mollica, and Evelina Fedorenko. 2024. Distributed Sensitivity to Syntax and Semantics throughout the Language Network. *Journal of Cognitive Neuroscience*, pages 1–43.

Sneha Shashidhara, Floortje S. Spronkers, and Yaara Erez. 2020. Individual-subject functional localization increases univariate activation but not multivariate pattern discriminability in the “multiple-demand” frontoparietal network. *Journal of Cognitive Neuroscience*, 32(7):1348–1368.

Michael Siegal and Rosemary Varley. 2006. Aphasia, language, and theory of mind. *Social Neuroscience*, 1(3–4):167–174.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

Haiyang Sun, Lin Zhao, Zihao Wu, Xiaohui Gao, Yutao Hu, Mengfei Zuo, Wei Zhang, Jun-Feng Han, Tianming Liu, and Xintao Hu. 2024. Brain-like functional organization within large language models.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. 2023. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu 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. *ArXiv*, abs/2307.09288.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024a. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024b. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, pages 1–18.

Rosemary Varley and Michael Siegal. 2000. Evidence for cognition without grammar from causal reasoning and ‘theory of mind’ in an agrammatic aphasic patient. *Current Biology*, 10(12):723–726.

Rosemary A. Varley, Nicolai J. C. Klessinger, Charles A. J. Romanowski, and Michael Siegal. 2005. Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9):3519–3524.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Net-*

works for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Alexandra Woolgar, Alice Parr, Rhodri Cusack, Russell Thompson, Ian Nimmo-Smith, Teresa Torralva, Maria Roca, Nagui Antoun, Facundo Manes, and John Duncan. 2010. Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 107(33):14899–14902.

Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. 2023. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust llm evaluators. *Preprint*, arXiv:2405.11966.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling a core linguistic region in large language models. *Preprint*, arXiv:2310.14928.

Appendix

A Functional Localizers

Figure 7 shows a pair of examples for each network localizer stimuli. We provide more details of each stimuli set below.

A.1 Language Localizer

The language localizer uses the same set of 240 sentences and 240 lists of non-words[2] as used by neuroscientists to localize the human language network. Each sentence consists of 12 words, and each list of non-words consists of 12 non-words to control for length. Since we are using a trained BPE tokenizer that breaks down each word into a number of tokens, we truncated the tokens to have a maximum length of 12 to ensure similar sequence length.

A.2 Multiple Demand Localizer

The arithmetic multiple-demand localizer used in neuroscience includes a set of "hard" arithmetic questions alongside a set of "easy" ones. These stimuli are usually generated by a MATLAB script that displays a mathematical problem on a screen for participants to solve, followed by two answer choices, one of which is correct. "Hard" questions are defined as addition problems with results exceeding 20 (e.g., 18+5), while "easy" questions yield results below 10 (e.g., 4+2). We adapted this localizer by similarly generating hard and easy arithmetic questions but slightly increased the complexity. Specifically, for hard questions, we sampled two numbers between 100 and 200, with each problem randomly chosen to be either an addition or subtraction with equal likelihood. For easy questions, we sampled two numbers in the range of 1 to 20. We generated 100 questions for each stimuli set. Examples are shown in Figure 7.

A.3 Theory of Mind Localizer

We use the same set of stimuli employed in neuroscience to localize the theory-of-mind network in the human brain (Dodell-Feder et al., 2011), which includes 10 false-belief stories and 10 false-photograph stories[3]. The prompt was structured to mirror the instructions given to participants during the neuroimaging study, followed by the story, the question, two answer choices (True or False), and an answer. Example of the prompt given from each set are shown in Figure 7. When evaluating the model on the test-set, we give it the following instruction: "The following multiple choice questions is based on the following story. The question is related to Theory-of-Mind. Read the story and then answer the questions. Choose the best answer from the options provided by printing it as is without any modifications."

B Localized Units Location

B.1 Language Units

Figure 8 shows the distribution of the top 1% language selective units for all 18 models tested in this work. An interesting observation is that the distribution of language-selective units remains similar in models both before and after instruction tuning.

B.2 Multiple Demand Units

Figure 9 shows the distribution of the top 1% Multiple Demand (MD) selective units for the 10 models tested for MD in this work.

B.3 Theory of Mind Units

Figure 10 shows the distribution of the top 1% Theory of Mind (ToM) selective units for the 10 models tested for ToM in this work. The ToM selective units are more distributed across the model layers rather than being more clustered as in MD and the language-selective units. This might be due to the small number of stimuli samples used for the ToM localizer.

C Models

Table 2 lists all 18 models analyzed in this study and indicates which models were used in which experiment. We kept 10 models for each experiment as shown in the last row. Table 11 shows the number of units corresponding to each percentage level for all models.

D Finegrained Results

D.1 Language Results

Tables 6 and 7 display results for the 10 models tested on three language benchmarks—SyntaxGym, BLiMP, and GLUE—along with the

[2] The language localizer stimuli were retrieved from: https://www.evlab.mit.edu/resources-all/download-localizer-tasks

[3] The Theory of Mind localizer stimuli were retrieved from https://saxelab.mit.edu/use-our-efficient-false-belief-localizer/

	Model	Lang	MD/ToM
	GPT2-Large	✓	✗
	GPT2-XL	✓	✗
	Gemma-2B	✓	✗
	Gemma-7B	✓	✗
	Gemma-1.1-7B-Instruct	✗	✓
	Phi-3.5-Mini-Instruct	✓	✓
	Mistral-7B-v0.3	✓	✗
	Mistral-7B-Instruct-v0.3	✗	✓
	LLaMA-2-7B	✓	✗
	LLaMA-2-7B-Chat	✗	✓
	LLaMA-2-13B	✓	✗
	LLaMA-2-13B-Chat	✗	✓
	LLaMA-3.1-8B-Instruct	✓	✓
	LLaMA-3.2-3B-Instruct	✗	✓
	Falcon-7B	✓	✗
	Falcon-7B-Instruct	✗	✓
	Qwen2.5-3B-Instruct	✗	✓
	Qwen2.5-7B-Instruct	✗	✓
#	18	10	10

Table 2: **Models Used in This Work** Overview of the 18 models analyzed, with an indication of the specific experiments in which each model was used. Lang referes to the language experiments, MD refers to the Multiple Demand experiments, and ToM refers to the Theory of Mind experiments.

average performance across these benchmarks. Each model's performance is shown initially without ablation, followed by ablations of the top-0.125, 0.25, 0.5, 1% language-selective units, and then with randomly sampled units at the same percentages. The performance changes in Figure 3 can be reproduced by subtracting post-ablation results from the baseline (0%) for both language-selective and random unit ablations. Results with random units are averaged across three random seeds.

D.2 Multiple Demand Results

Table 8 presents the results for the 10 models tested on the MATH benchmark, organized by difficulty level and including the overall macro average across levels. Each model's performance is shown under three conditions: without ablation, after ablating the top 1% of Multiple Demand-selective units, and with an equivalent number of randomly sampled units.

D.3 Theory of Mind Results

Table 9 similarly shows the results for the 10 models tested on the ToMi benchmark, organized by question type, whether it involves a false-belief scenario (n=231) or true-belief scenarios (n=389), and including the macro average across both types. Each model's performance is shown under three conditions: without ablation, after ablating the top 1% of theory-of-mind-selective units, and with an equivalent number of randomly sampled units. Table 10 shows the same but when ablating the top-2% of units.

E More Brain Alignment Statistical Tests

In Section 7, we performed Welch's t-test to demonstrate that units from the language network are significantly more brain-aligned than three sets of randomly sampled units from the model, particularly when sampling a small number of units. Here, we conduct the Shapiro-Wilk test to verify that each distribution follows a normal distribution, as Welch's t-test assumes normality in the compared distributions. Tables 3 and 4 present the test statistics and p-values for the brain alignment results across models, comparing both language and random units at each percentage and for each dataset. These results confirm that the distributions are indeed normal. A p-value greater than 0.05 indicates normality, while values below this threshold suggest deviation from normality. Notably, the only cases where the p-value falls below 0.05—indicating non-normal distributions—are for the 0.5% and 1% conditions in the Tuckute2024 dataset, where no significant difference was observed.

Percentage	Language Units	Random Units
0.03125%	(0.902, 0.233)	(0.971, 0.565)
0.0625%	(0.957, 0.755)	(0.939, 0.085)
0.125%	(0.955, 0.722)	(0.981, 0.853)
0.25%	(0.933, 0.475)	(0.962, 0.345)
0.5%	(0.945, 0.609)	(0.974, 0.658)
1%	(0.945, 0.612)	(0.970, 0.551)

Table 3: Shapiro-Wilk test (statistics and p-values) for brain alignment distributions across models. The test is conducted separately for language units and randomly sampled units at each percentage level for the Pereira2018 dataset.

We also conducted a permutation test, a non-parametric statistical method that does not re-

Percentage	Language Units	Random Units
0.03125%	(0.948, 0.641)	(0.9511, 0.181)
0.0625%	(0.962, 0.802)	(0.9507, 0.176)
0.125%	(0.973, 0.915)	(0.9810, 0.852)
0.25%	(0.914, 0.309)	(0.9592, 0.296)
0.5%	(0.829, 0.032)	(0.9693, 0.519)
1%	(0.825, 0.029)	(0.9725, 0.608)

Table 4: Shapiro-Wilk test (statistics and p-values) for brain alignment distributions across models. The test is conducted separately for language units and randomly sampled units at each percentage level for the TUCKUTE2024 dataset.

quire the assumption of normality. This method involves randomly shuffling data labels across 10,000 permutations to generate a null distribution of the test statistic. By comparing the observed test statistic to this null distribution, we evaluated the statistical significance of our results. The findings from the permutation test confirmed the significance of our results, as shown in Table 5.

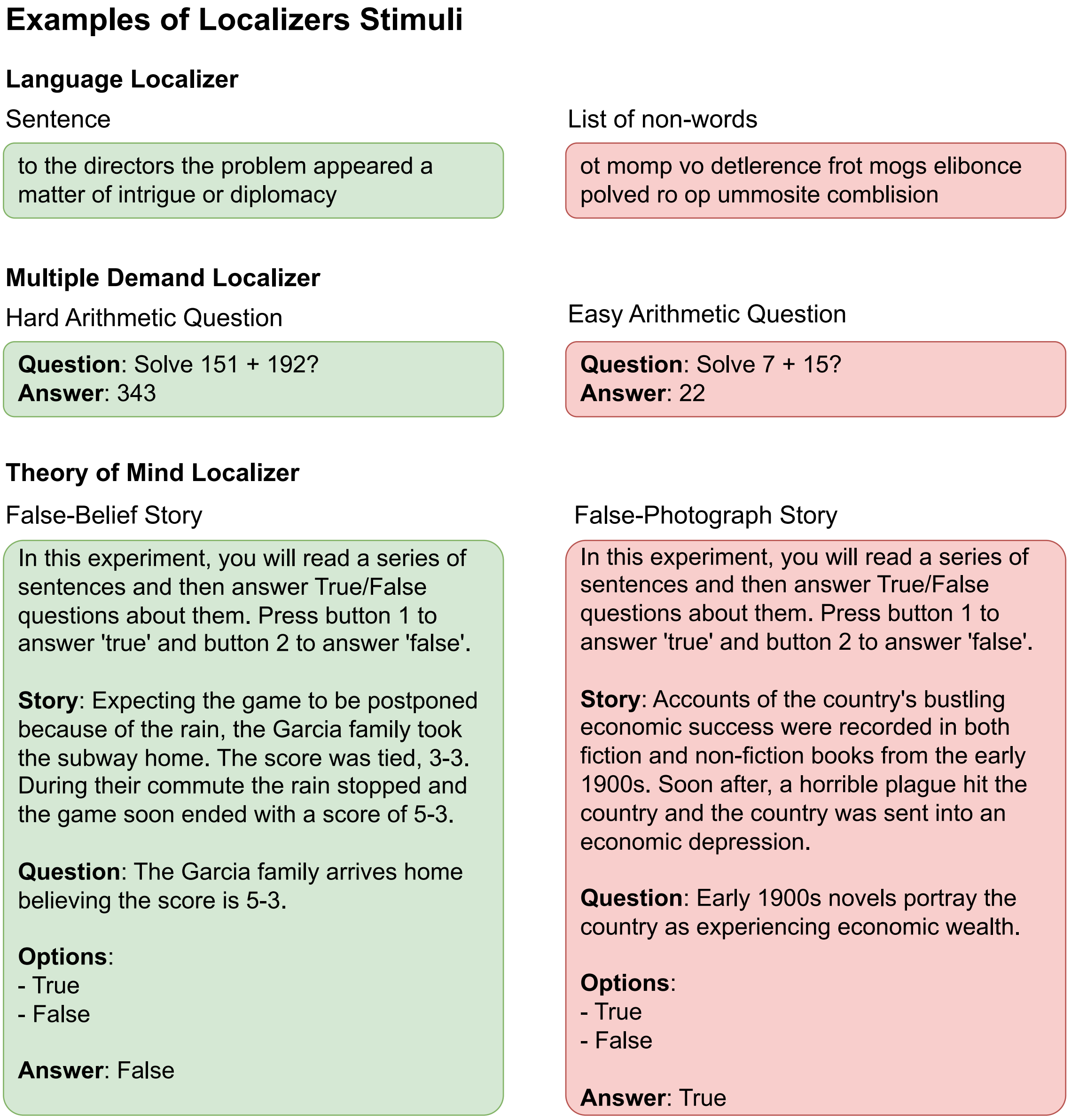
Percentage	PEREIRA2018	TUCKUTE2024
0.03125%	0.001	0.004
0.0625%	0.004	0.013
0.125%	0.138	0.000
0.25%	0.548	0.013
0.5%	0.195	0.169
1%	0.084	0.458

Table 5: Permutation test p-values assessing the statistical significance of brain alignment differences on both datasets. The test was conducted by randomly shuffling data labels across 10,000 permutations to generate a null distribution of the test statistic. The observed test statistic was then compared to this null distribution to compute the p-values. Lower p-values indicate stronger evidence against the null hypothesis, confirming the robustness of our findings.

F Brain-Score Datasets

Tuckute2024 This dataset consists of 5 participants reading 1000 6-word sentences presented one sentence at a time for 2s. BOLD responses from voxels in the language network were averaged within each participant and then across participants to yield an overall average language network response to each sentence. The stimuli used span a large part of the linguistic space, enabling model-brain comparisons across a wide range of single sentences. Sentence presentation order was randomized across participants. The averaging of sentences across participants effectively minimizes the effect of temporal autocorrelation/context in this dataset. In combination with the diversity in linguistic materials, this dataset presents a particularly challenging dataset for model evaluation. The noise ceiling for TUCKUTE2024 is $r = 0.56$, see Tuckute et al. (2024b) for more details.

Pereira2018 This dataset consists of fMRI activations (blood-oxygen-level-dependent; BOLD responses) recorded as participants read short passages presented one sentence at a time for 4 s. The dataset is composed of two distinct experiments: one with 9 subjects presented with 384 sentences, and another with 6 subjects presented with 243 sentences each. The passages in each experiment spanned 24 different topics. The results reported for this dataset are the average alignment across both experiments (Pereira et al., 2018). The reported results for this dataset use an unshuffled cross-validation scheme, ensuring that sentences from the same passage appear exclusively in either the training or testing set.



Examples of Localizers Stimuli

Language Localizer

Sentence

to the directors the problem appeared a matter of intrigue or diplomacy

List of non-words

ot momp vo detlerence frot mogs elibonce polved ro op ummosite comblision

Multiple Demand Localizer

Hard Arithmetic Question

Question: Solve 151 + 192?
Answer: 343

Easy Arithmetic Question

Question: Solve 7 + 15?
Answer: 22

Theory of Mind Localizer

False-Belief Story

In this experiment, you will read a series of sentences and then answer True/False questions about them. Press button 1 to answer 'true' and button 2 to answer 'false'.

Story: Expecting the game to be postponed because of the rain, the Garcia family took the subway home. The score was tied, 3-3. During their commute the rain stopped and the game soon ended with a score of 5-3.

Question: The Garcia family arrives home believing the score is 5-3.

Options:
- True
- False

Answer: False

False-Photograph Story

In this experiment, you will read a series of sentences and then answer True/False questions about them. Press button 1 to answer 'true' and button 2 to answer 'false'.

Story: Accounts of the country's bustling economic success were recorded in both fiction and non-fiction books from the early 1900s. Soon after, a horrible plague hit the country and the country was sent into an economic depression.

Question: Early 1900s novels portray the country as experiencing economic wealth.

Options:
- True
- False

Answer: True

Figure 7: **Examples of Localizers Stimuli**. Language stimuli consists of 240 sentences and 240 lists of non-words. Multiple Demand stimuli consists of 100 hard arithmetic problems and 100 easy ones. Theory of Mind consists of 10 false-belief stories and 10 false-photograph stories. The instruction given in the Theory of Mind stimuli is the same given to participants during the neuroimaging study. See Appendix A for more details about each localizer.

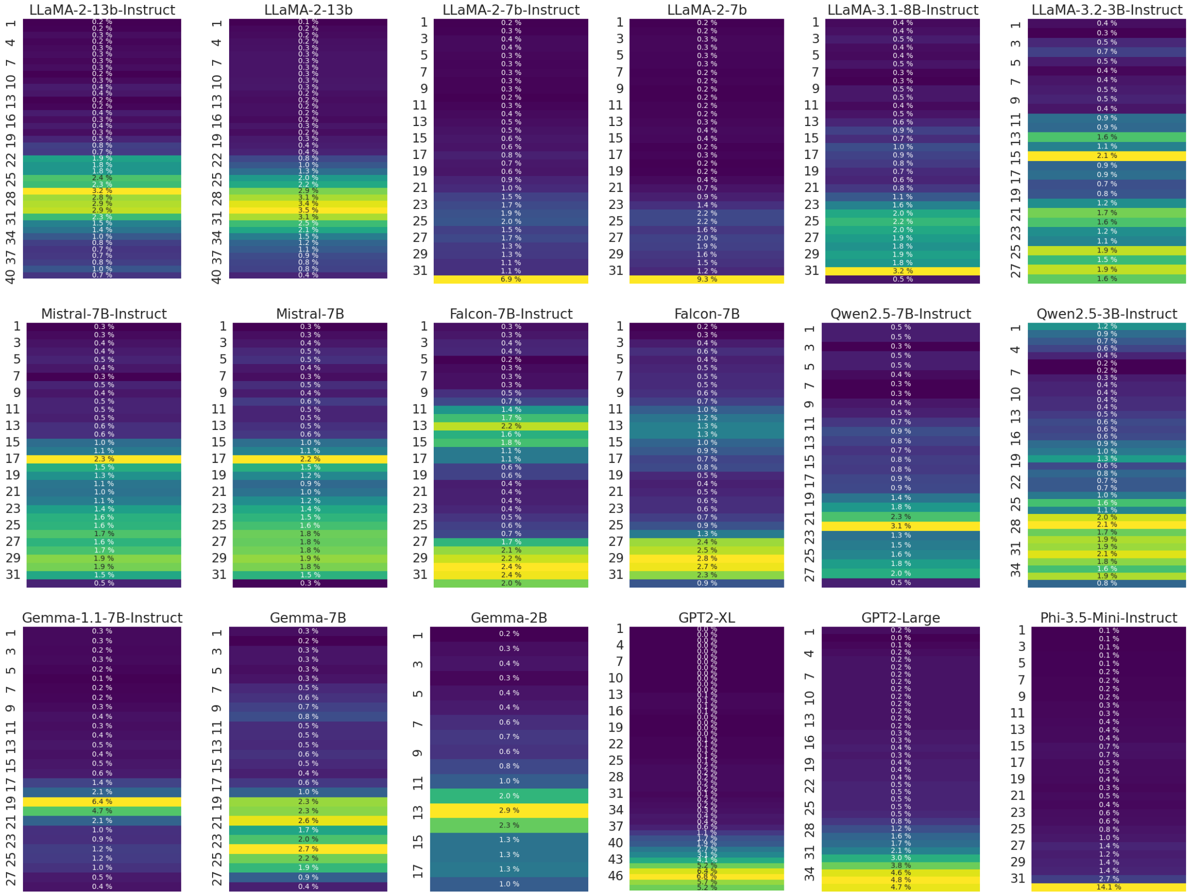


Figure 8: **Distribution of Language Units Across Layers**. The distribution of the top 1% most language-selective units across layers in all 18 models tested in this work. The models are displayed from top to bottom, with each layer labeled by the percentage of units identified as belonging to the top 1%.

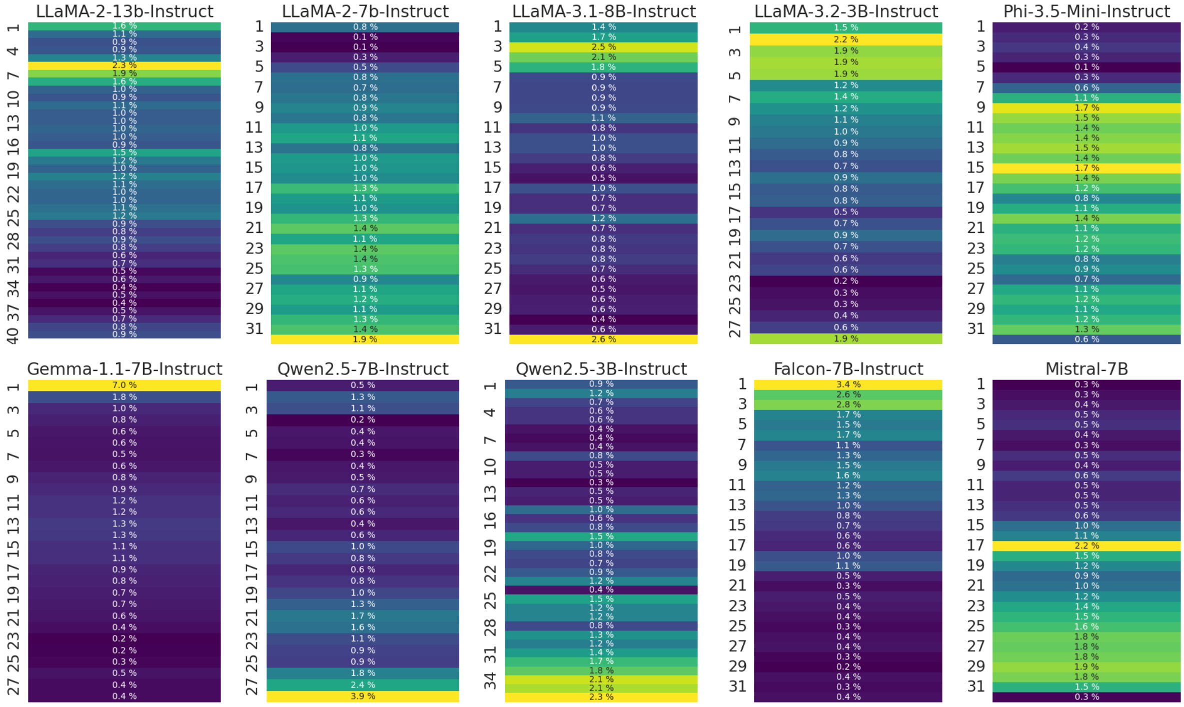


Figure 9: **Distribution of Multiple Demand Units Across Layers**. The distribution of the top 1% most Multiple Demand (MD) selective units across layers in the 10 models tested for MD in this work. The models are displayed from top to bottom, with each layer labeled by the percentage of units identified as belonging to the top 1%.

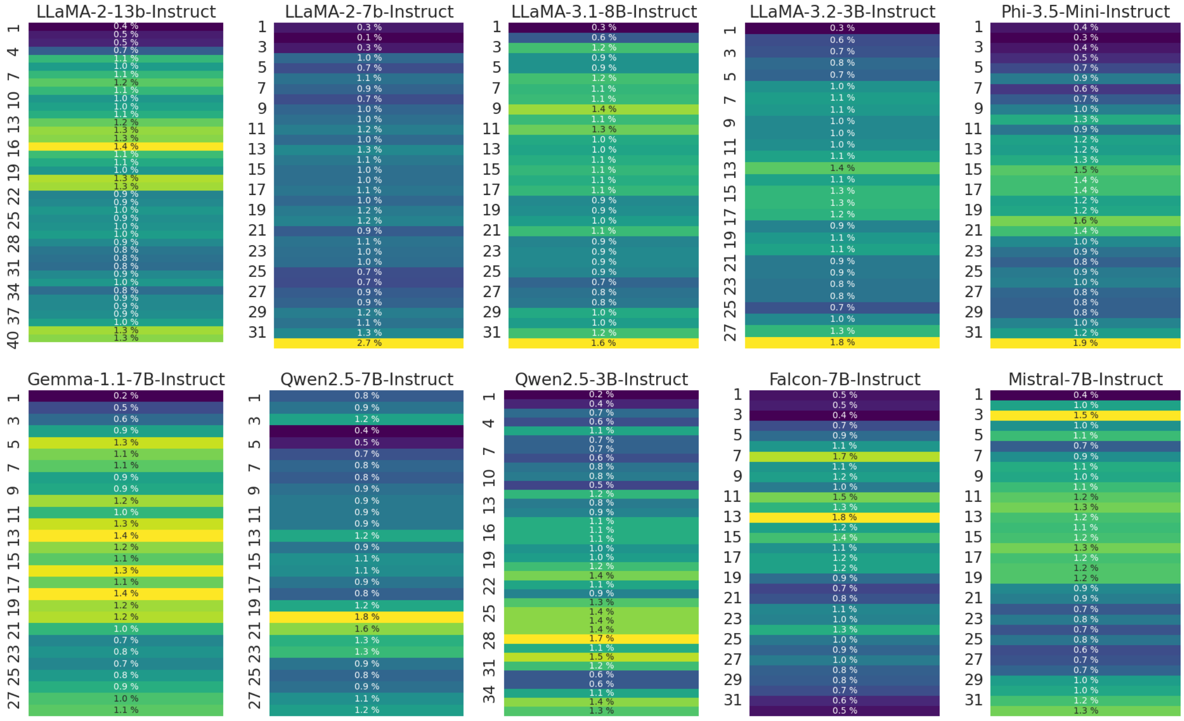


Figure 10: **Distribution of Theory of Mind Units Across Layers**. The distribution of the top 1% most theory of mind (ToM) selective units across layers in the 10 models tested for ToM in this work. The models are displayed from top to bottom, with each layer labeled by the percentage of units identified as belonging to the top 1%.

Model	Ablation Units	Percentage	SyntaxGym	BLiMP	GLUE	Average
GPT2-Large	-	0%	78.50	83.55	45.42	69.16
	Language	0.125%	61.87	76.79	44.23	60.96
	Language	0.25%	50.03	72.69	43.13	55.28
	Language	0.5%	46.99	69.37	38.55	51.64
	Language	1%	41.07	66.09	39.96	49.04
	Random	0.125%	78.02	83.50	45.39	68.97
	Random	0.25%	78.28	83.33	45.49	69.03
	Random	0.5%	77.95	82.89	44.48	68.44
	Random	1%	76.87	82.70	43.97	67.85
GPT2-XL	-	0%	82.70	83.38	46.85	70.98
	Language	0.125%	80.20	81.64	44.78	68.88
	Language	0.25%	70.73	78.06	46.24	65.01
	Language	0.5%	66.54	77.19	44.75	62.82
	Language	1%	56.02	74.86	43.12	58.00
	Random	0.125%	82.26	83.16	46.40	70.61
	Random	0.25%	80.76	83.07	46.38	70.07
	Random	0.5%	79.93	82.68	45.53	69.38
	Random	1%	79.44	81.64	45.19	68.76
Gemma-2B	-	0%	80.15	81.14	47.81	69.70
	Language	0.125%	38.16	56.34	41.79	45.43
	Language	0.25%	36.59	54.52	39.82	43.64
	Language	0.5%	26.02	52.54	37.38	38.64
	Language	1%	25.46	51.60	37.56	38.21
	Random	0.125%	80.18	81.10	47.35	69.54
	Random	0.25%	79.49	80.88	48.42	69.60
	Random	0.5%	79.51	80.93	46.25	68.90
	Random	1%	65.89	72.20	42.65	60.25
Gemma-7B	-	0%	80.37	81.75	62.29	74.80
	Language	0.125%	54.99	64.30	43.34	54.21
	Language	0.25%	52.91	61.17	44.38	52.82
	Language	0.5%	25.67	63.75	41.15	43.52
	Language	1%	29.61	48.97	45.90	41.50
	Random	0.125%	80.15	80.48	61.96	74.20
	Random	0.25%	80.44	81.24	60.59	74.09
	Random	0.5%	80.55	81.25	63.05	74.95
	Random	1%	79.65	79.98	58.36	72.66
Phi-3.5-Mini-Instruct	-	0%	81.86	80.63	70.73	77.74
	Language	0.125%	45.42	58.62	60.60	54.88
	Language	0.25%	34.81	55.72	49.65	46.72
	Language	0.5%	25.37	53.56	33.40	37.44
	Language	1%	22.90	53.79	46.26	40.98
	Random	0.125%	80.16	80.95	70.80	77.30
	Random	0.25%	81.83	81.64	69.64	77.70
	Random	0.5%	78.79	80.35	68.61	75.92
	Random	1%	79.80	79.05	69.19	76.01

Table 6: **Language Benchmarks Results 1** Results for the 5 models on the language benchmarks tested in this work. Random for each percentage is averaged across 3 seeds. The results when ablating random units is almost the same as ablating no units, while ablating language units lead to a sharp drop in performance. See Table 7 for the results of the other models.

Model	Ablation Units	Percentage	SyntaxGym	BLiMP	GLUE	Average
LLaMA-2-7b	-	0%	81.07	85.63	50.60	72.43
	Language	0.125%	46.07	66.85	41.91	51.61
	Language	0.25%	39.51	64.24	40.91	48.22
	Language	0.5%	28.86	57.07	32.57	39.50
	Language	1%	26.82	56.01	38.33	40.39
	Random	0.125%	81.09	85.57	50.74	72.47
	Random	0.25%	81.26	85.03	50.25	72.18
	Random	0.5%	80.23	84.68	51.15	72.02
	Random	1%	80.63	84.53	47.44	70.87
LLaMA-2-13b	-	0%	82.91	84.82	59.53	75.76
	Language	0.125%	78.57	81.38	48.05	69.33
	Language	0.25%	62.12	74.84	42.47	59.81
	Language	0.5%	23.85	51.23	29.16	34.75
	Language	1%	29.13	51.42	30.03	36.86
	Random	0.125%	82.43	84.79	58.76	75.33
	Random	0.25%	82.13	84.66	55.18	73.99
	Random	0.5%	82.06	83.77	57.52	74.45
	Random	1%	81.21	83.55	53.94	72.90
LLaMA-3.1-8B-Instruct	-	0%	80.05	81.90	69.20	77.05
	Language	0.125%	80.25	79.60	66.44	75.43
	Language	0.25%	78.22	76.96	61.43	72.20
	Language	0.5%	73.12	77.60	55.77	68.83
	Language	1%	54.12	67.17	46.36	55.88
	Random	0.125%	79.93	81.89	68.98	76.93
	Random	0.25%	79.99	81.88	68.71	76.86
	Random	0.5%	79.92	81.10	69.51	76.85
	Random	1%	79.14	81.73	67.41	76.09
Mistral-7B	-	0%	80.39	83.44	64.03	75.95
	Language	0.125%	70.08	75.38	47.33	64.26
	Language	0.25%	44.11	66.73	44.91	51.91
	Language	0.5%	37.60	66.39	43.74	49.24
	Language	1%	33.05	61.85	40.34	45.08
	Random	0.125%	80.28	83.34	63.54	75.72
	Random	0.25%	80.46	83.13	63.91	75.84
	Random	0.5%	80.34	82.62	62.75	75.24
	Random	1%	79.22	82.51	63.00	74.91
Falcon-7B	-	0%	80.05	80.35	48.36	69.59
	Language	0.125%	72.17	75.83	46.86	64.95
	Language	0.25%	69.99	71.67	47.23	62.97
	Language	0.5%	51.36	60.23	44.17	51.92
	Language	1%	25.79	55.42	45.11	42.10
	Random	0.125%	79.59	80.26	48.44	69.43
	Random	0.25%	79.89	80.35	48.83	69.69
	Random	0.5%	78.62	79.96	48.40	69.00
	Random	1%	78.32	79.99	48.85	69.05

Table 7: **Language Benchmarks Results 2** Results for 5 models on the language benchmarks tested in this work. Random for each percentage is averaged across 3 seeds. The results when ablating random units is almost the same as ablating no units, while ablating language units lead to a sharp drop in performance. See Table 6 for the results of the other models.

Model	Ablation	Level 1	Level 2	Level 3	Level 4	Level 5	Average
Phi-3.5-Mini-Instruct	-	52.33	41.50	40.45	35.11	31.91	40.26
	MD	38.37	32.31	33.90	28.44	27.33	32.07
	Random	49.92	43.50	41.52	37.31	35.71	41.59
LLaMA-3.1-8B-Instruct	-	40.00	37.41	36.23	36.36	39.21	37.84
	MD	37.21	34.13	33.18	33.61	40.45	35.72
	Random	36.20	35.15	33.96	35.95	38.59	35.97
Mistral-7B-Instruct	-	39.07	35.71	37.31	35.61	34.01	36.34
	MD	36.28	33.90	32.65	32.03	30.51	33.07
	Random	35.35	33.07	34.05	34.03	33.23	33.95
LLaMA-2-7b-Instruct	-	24.42	28.57	29.24	28.69	29.04	27.99
	MD	24.65	29.71	29.42	29.44	29.11	28.47
	Random	23.41	28.46	27.50	26.94	28.39	26.94
LLaMA-2-13b-Instruct	-	25.35	33.11	29.78	29.19	29.35	29.35
	MD	28.14	25.74	24.48	25.27	23.84	25.49
	Random	26.51	31.18	28.34	28.86	28.73	28.72
LLaMA-3.2-3B-Instruct	-	35.35	32.77	33.99	33.69	35.56	34.27
	MD	31.63	32.20	33.90	33.11	34.70	33.11
	Random	34.19	32.77	32.56	34.50	35.74	33.95
Gemma-1.1-7B-Instruct	-	40.00	37.41	35.96	34.28	35.87	36.71
	MD	35.58	34.81	30.49	31.28	32.53	32.94
	Random	37.21	36.28	35.19	33.75	35.12	35.51
Falcon-7B-Instruct	-	23.49	26.64	23.86	25.85	23.91	24.75
	MD	27.91	26.30	25.20	25.19	24.15	25.75
	Random	26.98	25.21	24.48	25.05	23.96	25.14
Qwen2.5-7B-Instruct	-	59.53	60.43	59.37	56.46	56.91	58.54
	MD	59.07	56.12	56.59	55.55	54.66	56.40
	Random	57.75	57.60	56.53	54.05	53.73	55.93
Qwen2.5-3B-Instruct	-	53.49	49.55	49.33	44.37	47.28	48.80
	MD	43.72	40.14	40.54	36.03	38.66	39.82
	Random	42.95	40.36	39.13	36.86	37.50	39.36

Table 8: **MATH Benchmark Results** Results for the 10 models tested on the MATH benchmark, showing performance in the following conditions for each model: without ablation, with ablation of the top 1% most MD-selective, and with randomly sampled. The results for *Random* is averaged across 3 seeds. MD stands for multiple demand.

Model	Ablation Units	False Belief	True Belief	Average
Phi-3.5-Mini-Instruct	-	50.65	86.38	68.51
	ToM	17.75	98.46	58.10
	Random	17.75	96.92	57.33
LLaMA-3.1-8B-Instruct	-	80.95	75.32	78.14
	ToM	64.50	75.32	69.91
	Random	76.62	68.47	72.54
Mistral-7B-Instruct	-	79.22	65.81	72.52
	ToM	64.07	61.95	63.01
	Random	69.55	68.47	69.01
LLaMA-2-7b-Instruct	-	23.81	79.69	51.75
	ToM	20.78	79.95	50.36
	Random	32.47	68.38	50.42
LLaMA-2-13b-Instruct	-	63.64	68.38	66.01
	ToM	49.35	69.92	59.64
	Random	59.88	58.01	58.95
LLaMA-3.2-3B-Instruct	-	9.96	92.80	51.38
	ToM	9.52	91.77	50.65
	Random	16.88	85.52	51.20
Gemma-1.1-7B-Instruct	-	78.79	65.55	72.17
	ToM	62.77	66.07	64.42
	Random	72.87	63.58	68.23
Falcon-7B-Instruct	-	50.22	46.02	48.12
	ToM	49.78	45.50	47.64
	Random	52.67	48.41	50.54
Qwen2.5-7B-Instruct	-	97.84	41.65	69.74
	ToM	93.51	44.73	69.12
	Random	92.35	46.62	69.48
Qwen2.5-3B-Instruct	-	81.82	59.38	70.60
	ToM	77.06	56.56	66.81
	Random	46.90	60.07	53.48

Table 9: **TOMi Benchmark Results (1% Lesion)** Results for the 10 models tested on the TOMi benchmark, showing performance in the following conditions for each model: without ablation, with ablation of the top 1% most ToM-selective, and with randomly sampled. The results for *Random* is averaged across 3 seeds. ToM stands for theory of mind.

Model	Ablation Units	False Belief	True Belief	Average
Phi-3.5-Mini-Instruct	-	50.65	86.38	68.51
	ToM	7.36	97.69	52.52
	Random	27.71	92.03	59.87
LLaMA-3.1-8B-Instruct	-	80.95	75.32	78.14
	ToM	49.35	64.52	56.94
	Random	61.62	56.56	59.09
Mistral-7B-Instruct	-	79.22	65.81	72.52
	ToM	40.26	71.98	56.12
	Random	66.67	66.50	66.58
LLaMA-2-7b-Instruct	-	23.81	79.69	51.75
	ToM	19.05	77.63	48.34
	Random	35.93	63.75	49.84
LLaMA-2-13b-Instruct	-	63.64	68.38	66.01
	ToM	42.42	60.15	51.29
	Random	52.53	54.07	53.30
LLaMA-3.2-3B-Instruct	-	9.96	92.80	51.38
	ToM	19.48	82.01	50.74
	Random	25.83	76.86	51.35
Gemma-1.1-7B-Instruct	-	78.79	65.55	72.17
	ToM	61.04	67.87	64.45
	Random	71.57	64.27	67.92
Falcon-7B-Instruct	-	50.22	46.02	48.12
	ToM	52.81	46.27	49.54
	Random	50.07	50.64	50.36
Qwen2.5-7B-Instruct	-	97.84	41.65	69.74
	ToM	89.61	50.13	69.87
	Random	60.03	59.13	59.58
Qwen2.5-3B-Instruct	-	81.82	59.38	70.60
	ToM	87.45	37.28	62.36
	Random	64.07	46.02	55.04

Table 10: **TOMi Benchmark Results (2% Lesion)** Results for the 10 models tested on the TOMi benchmark, showing performance in the following conditions for each model: without ablation, with ablation of the top 2% most ToM-selective, and with randomly sampled. The results for *Random* is averaged across 3 seeds. ToM stands for theory of mind.

Model	**0.03125%**	**0.0625%**	**0.125%**	**0.25%**	**0.5%**	**1%**	**2%**
Falcon-7B	45	90	181	363	727	1454	2908
Falcon-7B-Instruct	45	90	181	363	727	1454	2908
GPT2-Large	14	28	57	115	230	460	921
GPT2-XL	24	48	96	192	384	768	1536
Gemma-1.1-7B-Instruct	26	53	107	215	430	860	1720
Gemma-2B	11	23	46	92	184	368	737
Gemma-7B	26	53	107	215	430	860	1720
LLaMA-2-13b	64	128	256	512	1024	2048	4096
LLaMA-2-13b-Instruct	64	128	256	512	1024	2048	4096
LLaMA-2-7b	40	81	163	327	655	1310	2621
LLaMA-2-7b-Instruct	40	81	163	327	655	1310	2621
LLaMA-3.1-8B-Instruct	40	81	163	327	655	1310	2621
LLaMA-3.2-3B-Instruct	26	53	107	215	430	860	1720
Mistral-7B	40	81	163	327	655	1310	2621
Mistral-7B-Instruct	40	81	163	327	655	1310	2621
Phi-3.5-Mini-Instruct	30	61	122	245	491	983	1966
Qwen2.5-3B-Instruct	23	46	92	184	368	737	1474
Qwen2.5-7B-Instruct	31	62	125	250	501	1003	2007

Table 11: **Number of Units at Specified Percentage Levels for Each Model** The table shows the number of units corresponding to each percentage level (x%) for each model. These values are calculated by multiplying the number of layers, by the hidden dimension, and the specified percentage.