Mechanistic Interpretability of Grammatical Structures and Fuzzy Prompts

Chirag Adwani[*]
cadwani@umd.edu

Marvyn Bailly[*]
mbailly@umd.edu

Kejia Zhang[*]
zkj15@umd.edu

1 Introduction

Since their emergence, large language models (LLMs) have demonstrated remarkable abilities in understanding and generating human language across diverse linguistic systems. They can often interpret multiple languages - and even produce coherent responses to prompts that are grammatically irregular or incomplete. Yet, how LLMs internally represent and process linguistic structure remains an open and fundamental question.

Recent work by (Zhang et al., 2025) introduced tools such as path patching, adapted from mechanistic interpretability, along with logit attribution, to study cross-linguistic structural representations. Their study revealed intriguing similarities and divergences in internal circuitry across languages, suggesting that multilingual models may develop partially shared but also language-specific mechanisms for encoding grammar and syntax. Building on this foundation, our project seeks to reproduce and extend their results while broadening the methodological scope of mechanistic interpretability in the multilingual setting.

Specifically, we aim to examine whether multilingual LLMs employ shared or distinct internal circuits when processing languages with markedly different grammatical logics, such as English and Chinese. Chinese, spoken by the largest number of people worldwide, differs substantially from English in its syntactic and morphological organization, providing a natural testbed for probing cross-linguistic generalization. We will employ both path patching - to identify causal components within transformer architectures and the Information Flow Route (IFR) framework, which traces how information propagates through residual streams at each layer.

By constructing parallel English–Chinese datasets that isolate specific grammatical features, we will analyze whether comparable attention heads or feed-forward modules are activated across languages, and how these activation patterns change when grammatical cues are degraded through fuzzy or ungrammatical prompts. Through these experiments, we aim to deepen our understanding of how LLMs internalize grammatical structure, how such representations vary across languages, and whether universal computational principles underlie multilingual language understanding.

2 Related work

The primary inspiration for this project comes from Zhang et al. (2025) (Zhang et al., 2025), who conducted one of the first large-scale studies on mechanistic interpretability in multilingual language models. Their work introduced the use of path patching and logit attribution to analyze structural similarities and differences in how LLMs encode grammatical information across languages. Building on these techniques, we aim to reproduce and extend their findings by investigating whether similar interpretability tools can be applied to less structured or "fuzzy" linguistic inputs.

Path patching has emerged as a core method in mechanistic interpretability research (Wang et al., 2023). It seeks to identify the specific internal circuits such as attention heads and MLP neurons that causally drive particular model behaviors. The method compares a clean input (where the model performs correctly) with a corrupted input (where it fails), then systematically replaces or "patches" internal activations from the clean run into the corrupted one. If patching a specific component restores correct behavior, that com-

[*] Equal contribution.

ponent is deemed causally important. Through this process, researchers can map functional circuits responsible for reasoning, often supported by logit attribution, which projects activations into the model's output (or "verb") space to reveal their semantic effects (Belrose et al., 2023).

Although widely adopted, path patching exhibits key limitations. The discovered circuits often depend heavily on the design of minimal pair templates (e.g., clean vs. corrupted examples), which must be carefully constructed to isolate a single behavior. Such pairs, however, are difficult to generalize across languages, domains, or grammatical contexts, limiting the interpretive scope of the method.

In contrast, the Information Flow Route (IFR) framework (Ferrando and Voita, 2024) provides a more continuous and general approach. IFR measures how information propagates through residual streams and attention pathways at each layer, allowing researchers to quantify each component's contribution without relying on handcrafted input pairs. By tracing directed information flow, IFR captures the global structure of computation within the model and can be applied to a wider range of linguistic or conceptual settings.

In this project, we plan to first reproduce the cross-linguistic results of Zhang et al. (2025) to validate our implementation of these interpretability techniques. We will then extend their approach by applying path patching and IFR to study how LLMs respond to fuzzy or ungrammatical prompts, a setting that—despite the growing literature on mechanistic interpretability—has not yet been systematically explored.

3 Your approach

Following in the footsteps of (Zhang et al., 2025), we plan to carry out a controlled probe task known as Indirect Object Identification (Wang et al., 2023) in the hope to duplicate the found correlations between attention-head frequencies of English and Chinese prompts given to a multilingual LLM like BLOOM-560M (Workshop et al., 2023), or Qwen2-0.5B-instruct, and monolingual models such as GPT-2small and CPM-Generate. To better understand a specific natural language task, we will use indirect object identification. An example of an IOI prompt "Susan and Mary went to the bar. Susan gave a drink to [BLANK]". The correct answer here is, of course, "Mary". We also generate pairs of data that exhibit one grammatical feature (e.g. plurality) that is present in one language but not in the other. An example of such a sentence is "Alice owns noun_singular. Bob owns noun_plural". We could continue to study the effect of a fuzzy search by masking the grammatical words in the IOI prompts.

We plan to proceed in two phases: (1) reproduce IOI results on the multi- and monolingual LLM models in English and Chinese, including head-level activity maps and correlation measurements. (2) Test whether heads implicated in grammatical processing remain necessary when prompts are made with typos or just as a bag-of-words ("fuzzy"). We can conclude by quantifying the cross-lingual circuit overlaps.

baselines To determine if LLMs use different internal circuits to handle processes with different grammatical structures, we will use IOI statements that contain a specific grammatical feature. We will use both path patching and IFR to analyze which heads and layers play significant roles in answering the statements. Additionally, we will use the Pearson Correlation Coefficient ρ to compute the similarity between activation paths (Freedman et al., 2007). An analysis of the similar pathways will be carried out to determine how similar the pathway flows are and if the models active similar heads.

To investigate the absence of grammar in prompt, we will use path patching and IFR to analyze the significant heads. We will investigate the similarities across models and study the effects of ablating significant heads or feed forward layers. We will expect similar results as shown in (Zhang et al., 2025) when recreating results.

To (in)validate our hypotheses, a similar method of analyses will preformed on different grammatical structures and fuzzy prompts.

compute resource justification The main computational cost of our project will be preforming the mechanistic interpretability of our models. We plan to use a dataset of around 900 pairs. To preform path patching, we will require two runs per pair. Thus to run a pass on a model with 12 layers and 12 heads, we would require $900 \times 2 \times 12 \times 12 = 270,000$ forward passes. A group member has a personal RTX 4070 to preform the computations on noting that the intended tools are required to be extendable to the GPU. Due to the size of the models and the dataset, the information should fit on the RAM of the GPU. The tasks can be easily parallelized since the 900 runs of pairs of path patching are embarrassingly parallel. If the computation time overhead provides an issue, we can use Colab.

3.1 Schedule

We plan to break down our project into the following list of sub tasks. We give an estimate for how long each task will take, noting that the total time adds up to 2.5 months

1. Read and learn more about Patch Patching, information flow route methods, and the tools needed to implement these methods like TransformerLens (Nanda and Bloom, 2022). (2 weeks)
2. Implement the required codes, run preliminary tests. (3 weeks)
3. Use the working model to recreate results obtained in (Zhang et al., 2025). Analyze. (2 weeks)
4. Create new results for proposed fuzzy search experiments and/or new grammatical pattern. Analyze. (2 weeks)
5. Work on final reports. (1 week)

Although computations will be carried out on a single members GPU, the group plans to work together to write code, analyze the results, create meaningful plots, and produce the final write up.

4 Data

The dataset will contain $60 \times 15 = 900$ sequences of IOI sentences generated using 15 different templates across 60 input words. The templates will be designed manually along with a Python script that, given an input word, fills in the template sentences. We will use ChatGPT to translate the templates and names into Chinese and have them manually inspected by a native speaker before adding them to the dataset. Note that the Python script can be easily modified to create pairs to be used in path patching. We can derive fuzzy variants of the said pairs by removing function words, by shuffling the order, or by introducing random incorrect spellings.

To generate a list of words to feed into our templates, we plan to utilize Hugging Face's database which we can curl for free. From the database of words (e.g. verbs or plural nouns) we can sort them by words which follow the exact grammatical pattern to be studied (e.g. suffix is "ed" or "s"). We will continue to translate the words into Chinese using ChatGPT and have them manually inspected by a native speaker.

This pipeline will allow us to generate 60 words that follow a grammatical pattern, translate them into Chinese, and finally create 15 sample inputs per word in English and Chinese using a Python script. This process should be fairly robust and in the case that Hugging Face's database does not contain the desired dataset, we can manually generate a set of 60 words using ChatGPT and manually fairing their quality. Furthermore, if the dataset proves inefficient, this method allows us to quickly generate another set.

5 Tools

The framework of this study will be implemented using Python to leverage the number of existing machine learning libraries. We will use the pretrained multilingual language models BLOOM-560M and Qwen2-0.5B-instruct as well the monolingual modals GPT-2 small (English) and CPM-Generate (Chinese). These models can be accessed through the Python library Transformers provided by Hugging Face. We note that the library can be easily extended to the GPU and parallelized.

To perform mechanistic interpretability to analyze the previously presented models, we utilize the libraries PyTorch and TransformersLens. TransformerLens provides tools for preforming path patching and ablating individual attention heads or feed-forward layers. We plan to manually implement the flow routes method as it is a newer method.

6 AI Disclosure

- It should be noted that the following AI tools were used in the creation of this proposal:
  - Gemini
  - ChatGPT

- Experiences

  Marvyn:

- My experience with Gemini was overall beneficial and reduced the time in the ”researching” phase of writing this proposal. I used the model to find tools such as Hugging Face which will provide access to the different LLM versions or where to find databases.
- I also asked Gemini some clarifying questions about the reference papers. Such as ”how did this paper generate their sample data?”

Chirag:

- I used ChatGPT to summarise the relevant papers, and I think it helped to reduced the time to go through them by a lot. It also helped me discover the other relevant works and understand the history of Mech Interp as a growing subject. I had a good conversation with ChatGPT, which definitely made me more interested in the subject.
- I also used ChatGPT to write me some quick example codes of path patching and IFR so that I could get a gist of what they were. I think it was able to help me understand it to the level of writing this proposal.
- I also did some paraphrasing using ChatGPT (including this sentence) for better brevity.

Kejia Zhang:

- I used ChatGPT to convert my Mandarin and English mixed version sentences to english.
- I used ChatGPT to check the fluency of my texts.
- My overall experience is positive. I usually need to change some specific misunderstanding words with chatGPT.

References

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Ferrando, J. and Voita, E. (2024). Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. W. W. Norton & Company, New York, 4th international student edition edition.

Nanda, N. and Bloom, J. (2022). Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2023). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L. V., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M. D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Zhang, R., Yu, Q., Zang, M., Eickhoff, C., and Pavlick, E. (2025). The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.