Statement of Purpose

of Prasoon Bajpai

My research interest lies in the interpretation of internal mechanisms of LLMs through mathematical modeling. I am also interested in the possibility of selectively editing model architecture to enhance and explain emergent behaviors at all parameter scales. I want to explore multimodal LLMs as reasoning engines. Particularly, I want to interpret their fused modality spaces and study their internal mechanisms to better guide their scaling methodologies leading to better models. I wish to design dynamic and high-bandwidth *thought data structures* to explain internal reasoning mechanisms of LLMs, irrespective of the modality present in the query. These can be used to create an internal fine-grained reasoning evaluation framework not focusing entirely on datasets and tasks.

The invention of language enabled effective communication channels for receiving feedback and spreading knowledge. The understanding of language has been closely linked with other neural activities in humans. I am highly interested in adopting principles of biological evolution to expand capabilities of LLMs. Given a fixed budget of parameter space, an attempt should be made to design more nuanced communication networks between smaller language models to reason effectively as a team. Therefore, believe there is scope of improving the gating mechanisms and overall communication layout in the Mixture-of-Experts (MoE) model methodology.

Research Work: I have been grateful to work as an NLP Graduate Researcher at Laboratory for Social Computation Systems under the guidance of Prof. Tanmoy Chakraborty. There, I had the opportunity to understand the general principles of conducting research while adapting and producing meaningful research volume in an accelerating field such as NLP. The technical discussions with people working in different areas not only kept me updated on latest research boundaries but also motivated me to come up my own research ideas that later translated to publications.

Long Context Capabilities in Multilingual Models: - (*Empirical Analysis*) Multilingual large language models (MLLMs) have demonstrated remarkable abilities in responding to queries in diverse languages [3]. In this work, I explored the multilingual **long-context handling capabilities** of open-source multilingual LLMs. We introduced a test to assess the models' ability to retrieve relevant information (*needle*) from a collection of multilingual distractor documents (*haystack*). I tested the multilingual question-answering capabilities across control parameters such as the language of the *needle* and the *haystack*, the position of the *needle* and the proposed context lengths of the models used. I found that MLLMs perform poorly in retrieving information from non-Latin *needles* irrespective of the language of the *haystack*. Moreover, the language of the distractor documents (*haystack*) did not affect MLLMs' performance at all. Furthermore, although some models claimed a context size of $8k$ tokens or greater, none demonstrated satisfactory cross-lingual retrieval performance as the context length increased. This work was **submitted to AAAI 2025** and is under review.

LLMs as Science Communicators: - (*Resources and Evaluation*) Motivated by cases of scientific misreporting of LLMs [1], I **curated a dataset** called `SCiPS-QA` which is a collection of boolean problems grounded on complex scientific objects, for assessing open-source and proprietary models on scientific question answering tasks that require a nuanced understanding and awareness of answer-ability. While most open-access models significantly underperformed compared to GPT-4 Turbo, myy experiments identified Llama-3-70B as a strong competitor, often surpassing GPT-4 Turbo in various evaluation aspects. I also found that the GPT models exhibit general incompetence in reliably verifying their responses. Moreover, I observed an alarming trend where human evaluators were deceived by incorrect responses from GPT-4 Turbo. This work was **accepted at EMNLP 2024 Main**.

Effect of Popularity on Internal LLM Structures - *(Mechanistic Interpretation)*: The positive correlation of string frequency in pertaining corpus to its retention in the parametric memory of LLMs is well established [4]. Motivated by psychological studies [5] on the difficulty of queries on task-addressing capabilities on humans, I created an experiment environment to scope in on the effect of popularity on the internal reasoning and knowledge-retrieval frameworks of LLMs. Particularly, I was interested in the effect of lexical variability of queries on the internal states of LLMs with the changing popularity of subjects in the queries. I found that LLMs address the relevant tokens of queries with higher popularity poorly compared to those of lower popularity. Even though the knowledge retrieval capabilities are shown to be higher when addressing higher popularity questions, the variability of the knowledge retrieved with the linguistic variations of query increases with the popularity of questions. This work has been **submitted to the Nature Communications Journal** and is under review.

Future Research Ideas - Following are some research directions that I would definitely want to explore.
Large Thought Models - *(Why must LLMs make Language Before Decision)*: Modelling *thoughts* as dynamic data structures which not only provide a high-bandwidth medium of information exchange but also allow for evaluating reasoning based capabilities of LLMs. This can be followed be designing more nuanced pre-training objectives to use transformers as overall *thought-modellers* and not just limited to modelling language.

Hallucination - *(Selectively Editing LLM Architecture for Mitigation)*: Hallucination defined as '*text generated which is unfaithful to a provided source input*' seems an insufficient definition to capture the illusory aspect of the text, especially in a reasoning argument. A metric-based definition of hallucination evaluating different internal (probability distribution of tokens) and external (faithfulness with the task) characteristics of generated text needs to be worked upon. I would like to study which structures of the transformer architecture increase exposure to hallucination risk. This would be followed by an attempt to mitigate such risk without limiting text characteristics such as creativity or diversity.

Multilingual LLMs - *(Interpreting Common Representation Spaces)*: A common internal representation of knowledge agnostic to languages should give rise to more emergent behaviours that are currently known. A study [2] found that multilingualism is positively associated with creativity. This research suggests that multilingual individuals develop richer cognitive networks due to their exposure to diverse linguistic structures. A parallel can be studied in multilingual LLMs. Moreover, multilingual individuals often report feeling like they *have different personalities* when speaking different languages. It will be interesting to see the existence of such behaviours in the context of multilingual LLMs.

Why PhD? I have been grateful to be surrounded by researchers of such calibre and dedication at LCS2. Being in a community of people, who share the same amount of passion for such an accelerated field, I never felt left behind. Prof. Tanmoy has had a particularly significant impact on me. He has always motivated to push myself beyond my boundaries. Under his guidance, I have overcome many obstacles to finally becoming super confident in my capabilities as a researcher. I believe that being a part of academia gives an individual the opportunity to foster progress further away in the future through interactions with future researchers. In the future, I would love to be a part of this community.

References

[1] Subhabrata Dutta and Tanmoy Chakraborty. "Thus Spake ChatGPT". In: *Commun. ACM* 66.12 (Nov. 2023), pp. 16–19. ISSN: 0001-0782. DOI: 10.1145/3616863. URL: https://doi.org/10.1145/3616863.

[2] Guillaume Fürst and François Grin. "Multicultural experience and multilingualism as predictors of creativity". In: *International Journal of Bilingualism* 25.5 (2021), pp. 1486–1494. DOI: 10.1177/13670069211019468. eprint: https://doi.org/10.1177/13670069211019468. URL: https://doi.org/10.1177/13670069211019468.

[3] Patrick S. H. Lewis et al. "MLQA: Evaluating Cross-lingual Extractive Question Answering". In: *CoRR* abs/1910.07475 (2019). arXiv: 1910.07475. URL: http://arxiv.org/abs/1910.07475.

[4] Alex Mallen et al. *When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.* 2023. arXiv: 2212.10511 [cs.CL]. URL: https://arxiv.org/abs/2212.10511.

[5] J. Elizabeth Richey et al. “More confusion and frustration, better learning: The impact of erroneous examples”. In: *Computers Education* 139 (2019), pp. 173–190. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2019.05.012. URL: https://www.sciencedirect.com/science/article/pii/S0360131519301277.