Request For Proposal

Project name:	**Memory-augmented LLMs**		
Complexity level	★★★★★	**Round 3 Pool**	S
Max award per proposal	$150,000 in AGIX	**Max amount for all proposals to be awarded in this round**	$250,000
Expected lead time	3-6 months after being awarded	**Technical requirements**	See below

Project description

The grand goal is to create language models, which can perform read and write operations over external memory in such a way that they memorize factual information not by adjusting their weights by gradient descent, but put it into the memory in one-shot manner. The primary objective now is to create PoC solutions based on the existing LLMs extended with external memory modules, which show capabilities of learning to memorize and retrieve relevant information from long contexts (10x longer than contexts of LLMs themselves).

Technical approach and available tools

The existing approaches to memory-augmented LLMs typically use a non-trainable memory with hand-coded algorithms to put information into memory and retrieve it. This either leads to memorizing everything in the extended context, which is still limited, and this adds a burden for the retrieval part, or to memorizing only certain types of facts (e.g., extracted subject-predicate- object triples). LLMs, in turn, learn important information, but implicitly via gradient descent. We propose to try training a memory model augmenting an LLM, which will not be trained to reproduce information from training sets directly, but will be trained to memorize and recall any new information. To avoid training from scratch, we propose to take a pre-trained SOTA foundation LLM.

Technical requirements

- API calls to external LLMs like ChatGPT are not used.
- Contemporary open-sourced LLMs (like LLAMA-2, but not necessarily it) are used.
- Memory is queried for read/write operations not via prompts, although a prompt-based solution can be used as baseline.
- The foundation LLM is not fine-tuned or trained by itself for memory querying, but an additional module inserted into it is trained entirely to perform read/write operations. However, the foundation LLM can be fine-tuned together with this module if needed. The module can be implemented as additional attention heads (not self-attention) on each or some levels of transformers, as hypernets influencing attention heads in the foundation model, as sampler controllers, or somehow else.
- External memory can be an embedding-based memory, knowledge graph, or something else as long as the requirements are satisfied.
- Computation time should not be more than 2x higher than that of the chosen base LLM.
- Memory work should be tested by feeding the model with text pieces of limited size (one piece per one inference run) alternated by questions about both recent and older pieces without providing additional context in question prompts themselves.
- Model and training code and checkpoints should be open sourced and the results should be reproducible.

Main proposal requirements

- Short summary.
- Requested total funding amount.
- Detailed description.
- Overview of team members and main experience.
- List of milestones, with per milestone, a demonstrable deliverable and required budget.

Assessment criteria

The main criteria our Technical team will be focusing on:

- Demonstration of training of memory r/w operations, e.g. the learning curve on data irrelevant to the training set.
- 75% recall rate on texts or dialogues up to 100–150 thousand tokens, which are read by the model piece-by-piece.

- Not degraded performance of the modified LLM on other standard benchmarks.
- Non-trainable memory, which meets the criteria, can be considered as a partial success.

Please try to be as objective as possible in estimating the value of your proposal. Only proposals of a high level of sophistication and quality can qualify for the maximum award.

In addition to this, the proposal will need to comply with all regular Deep Funding rules.

Please refer to the The General Rules And Conditions of Deep Funding portal for details on the proposal submission and further processes around the SNET RFP Pool.