FINAL RESEARCH REPORT

All Funded Projects 2024

MIT CSAIL Alliances

Project Title: Accurate and up-to-date language models with editable external memory

PI: Jacob Andreas

Project Summary

When the world changes, so does the text that humans write about it. How do we build language models that can be easily updated to reflect these changes? One popular approach is retrieval-augmented generation, in which new documents are inserted into a knowledge base and retrieved during prediction for downstream tasks. Most prior work on these systems have focused on improving behavior during prediction through better retrieval or reasoning. We introduce ERASE, which instead improves model behavior when new documents are acquired, by incrementally deleting or rewriting other entries in the knowledge base each time a document is added. In two new benchmark datasets evaluating models' ability to answer questions about a stream of news articles or conversations, ERASE improves accuracy relative to conventional retrieval-augmented generation by 7-13% (Mixtral-8x7B) and 6-10% (Llama-3-8B) absolute.

Problem Being Addressed

The world—and the language we used to describe it—are constantly changing. After reading the article "After Queen Elizabeth II died, the Queen's oldest son Charles has now become King Charles III", a knowledgeable reader might update an entire system of related beliefs, e.g., that King Charles III is now also the new head of Scotland. How can we train language models and other software systems to reflect these changes?

In language models, a simple (and often extremely effective) approach simply presents new information in models' inputs by leveraging either long-context methods (Tay et al., 2022) or retrieval augmented generation (RAG; Lewis et al., 2020a) which appends new documents to a knowledge base and retrieves a subset of relevant documents to condition on at prediction time. An important limitation of current RAG approaches is that they sometimes retrieve stale documents that have been invalidated by new information. We describe a method for retrieval-augmented generation that attempts to ensure that the external knowledge base always represents the current state of the world.

Project Status/Papers Published

Paper accepted to EMNLP Findings. Our method (ERASE) outperforms standard RAG baselines and long-context models, on a factual QA domain and single-hop section of a conversation domain. On the multi-hop subset, we find that ERASE performs comparably to baselines, suggesting there is room for future work to improve multi-hop memory editing

Project Title: Accurate and up-to-date language models with editable external memory (continued from above)

EMNLP Findings (see https://arxiv.org/pdf/2406.11830)

Student Funding

The funding paid for compute rather than Ras

Source Code

https://github.com/belindal/erase

Project Title: Provably secure LLMs that detect leaks with real-world evaluations

PI: Boris Katz / Andrei Barbu

Project Summary

Create a new tool which enables LLMs to operate in secure environments without the possibility of leaks. Ideally, we would want an LLM to have access to no more information than the user of that LLM. That way, no matter what the users demands of the LLM, they cannot learn more than they should know. Practically, this would require training an LLM for every user, which is not doable today and will not come to pass in the future.
Instead, we demonstrate new methods to merge LLM fine-tunings. All existing methods fail. We develop new datasets for compositional document QA (where you have multiple document silos, a user has access to a subset of them, and poses questions that require cross-silo inference and reasoning) as well as for leak detection (where you have multiple document silos and must determine if a statement was informed by some collection of those silos or not). Our method gives LLMs provable security, relying only on the bit-level permissions that operating systems already provide, enabling their deployments into the most sensitive environments.

Problem Being Addressed

Industry wants to give their employees and customers access to LLMs that know as much as possible and are as helpful as possible. But you don't want to give everyone access to everything on an honor code basis. Asking an LLM to not reveal some information or attempting to redact it after the fact has proven to be futile. There are countless workarounds to expose information.
We show how to synthesize an LLM at runtime that knows exactly what a user should be allowed to know and no more.
A related problem in industry is that even well meaning employees can leak information. There is no reliable way to detect leaks today. We developed the first dataset and model for detecting leaks in large organizations with LLMs.

Project Status

We submitted three papers for review, one of which is on arxiv; the others will be uploaded to arxiv soon. MIT patented the method.

We learned a tremendous amount about the problem during this process. For example, we thought that leak detection would be simple, but we quickly determined to our surprise that no appropriate leak detection datasets for testing AI/ML systems exist! Developing such datasets was a serious challenge. Similarly with document QA from multiple silos. In addition, we were

surprised by how brittle existing methods were.

We had hoped to adapt existing model merging methods or existing fine-tuning merging methods given the high performance they show in other domains. But those other domains essentially have information leakage where performance on one set of tasks helps performance on an unrelated set of tasks. In security data is orthogonal, knowledge about one set of secrets doesn't help you with another set of secrets. Once we developed an orthogonal setting and dataset, we found that all existing methods fail entirely. This forced us to develop new methods of combining fine tunings, which is an important outcome of the project.

Papers Published

Our papers are still in review.

Students Funded

MLA helped fund A. Alabdulakreem (master's student at the time and now incoming PhD student) and Vignesh Subramaniam (PhD student)

Grant Proposals

This research formed a core part of our proposal for the Air Force AI accelerator.

Source Code

https://github.com/Scuwr/SecureLLM

Project Title: Combining Symbolic Logic with Transformers for Interpretable and Controllable LLMs

PI: Yoon Kim

Project Summary

Despite the impressive capabilities of large language models (LLMs) based on the transformer architecture, they remain difficult to interpret and control. This is in contrast to classic symbolic approaches to NLP which provide a symbolic interface with which to interpret and control model behavior. However, symbolic systems are often inflexible and have difficulty processing inputs that do not conform to the symbolic specifications. This project seeks to improve Transformer-based language models through symbolic techniques.

Problem Being Addressed

Controlling and enhancing capabilities of Transformers through symbolic techniques

Project Status

Our first attempts at using first-order logic to enhance Transformers resulted in negative results. We are now currently working on learning a symbolic system (as latent variables) end-to-end, and am seeing some success.

Student Support

1

Project Title: Kodless: Building Entire Apps with LLMs

PI: Daniel Jackson

Project Summary

From our original proposal:

The goal of our project is to develop an approach for generating the code of an entire working application automatically, without requiring the developer to write any code, or even to review the code that is generated. The project will provide an embodiment of this approach in a tool that runs as a web service (so that applications can be generated without any installation of software). It will be evaluated in a series of case studies, and by a user test to calibrate how easy it is for developers to use the tool.

Problem Being Addressed

LLM-based code generation is advancing rapidly, with LLMs showing the ability to generate longer and more complex code fragments. But fundamentally LLM code generation is still focused on producing implementations of ad hoc components, that must be carefully defined by an engineer, and fit within a complicated and often brittle system context.

Most efforts at automating code generation are focused on AI agents that mimic human developers: playing the role of coding assistants to engineers, producing code fragments, test cases, helping suggest interfaces, and so on—but not changing the nature of programming in any fundamental way. In particular, these approaches can at best improve the productivity of engineers, but cannot extend the ability to create apps to end users (however well they may understand their problem domain). (There are also tools that generate entire apps, but their capabilities are very limited, and they tend to soon hit a complexity wall as requirements are added.) To fulfill the vision of a true app-generating agent that can be used by non-programmers, we believe that a new paradigm is needed that provides a new kind of software structure that is matched to the capabilities of LLMs.

Project Status

At the start of this project, we had already explored the generation of entire apps using concept structuring. Our prototype tool was demonstrated by generating a near functionally complete clone of Hacker News. This work did not, however, address the critical problem of *adaptability*: that all software evolves and must be adapted, even during the initial phases of development, in which features are being added, removed or modified.

In traditional LLM-based code generation, each modification of the code base cannot be independently verified as sound (that is, correctly fulfilling its intended function), and worse,

threatens the integrity of previously generated code. In short, modifications often don't work and risk breaking existing code.

We have addressed this challenge by developing a language of "synchronizations" that allows adaptations to be translated directly from individual user requirements to small and independent rules that determine how concepts are composed. We built an engine for executing synchronizations, and a set of prompts for generating both concepts and synchronizations from minimal application requirements. We have demonstrated this on the RealWorld benchmark, generating an entire backend for a Medium clone. In addition to offering an incremental and sound way to generate code, the improved modularity of our approach also pays off in greater succinctness, and our generated code is half the size of the comparable benchmark implementations that were produced by hand.

Papers/Conferences

We are working on a submission to a flagship software engineering and programming languages conference, due in March 2025.

Since the start of the grant (in September 2024), the PI (Daniel Jackson) has presented this work in invited talks at Brown University; at the New Directions in Software Technology (NDIST25) workshop; to the CSAIL Alliances (Byte Bites lecture series); and at UC Berkeley.

Student Funding

One PhD student and one UROP student have been funded on this grant.

Source Code

The code of our RealWorld implementation is available in GitHub (in a private repository) at https://github.com/eagonmeng/sync-realworld. Development of our engine, sync language, and LLM integration is constantly evolving, and we're excited to work with potential collaborators or end-users and give access to the repo to interested parties. Our sponsors are of course also welcome to peruse the repo.

The book by PI Daniel Jackson on concept-based development has its own website (https://essenceofsoftware.com) which contains blog articles, tutorials and more. The work described in this report has not yet been documented there, but we are planning a dedicated website for this specific project.

Chinese and Korean translations of the book were published last year; there was an earlier translation to Japanese and an upcoming translation to Russian to be published this year.

Concept development was previously adopted by Palantir (https://groups.csail.mit.edu/sdg/pubs/2023/concepts-onward-23.pdf), and last year was adopted by Autodesk (https://www.linkedin.com/posts/ogoldman_softwarearchitecturepractice-activity-7274830745548787712-vaES). The Axim Collaborative (who run the Open edX platform) commissioned the PI to conduct a concept analysis of their platform and build an initial version of a concept catalog, which is publicly available (https://publish.obsidian.md/axim).

Comments

In addition to making progress with this project, the PI (Daniel Jackson) has also been engaging with companies to adopt concept design as a general development strategy. This is important because the value of LLM-based code generation with concepts will be even higher if companies are already structuring software with concepts for strategic reasons. So far, Palantir and Autodesk have adopted concepts as central to their development process, and the Axim Collaborative (owner of the Open edX platform) has commissioned Jackson to conduct a concept analysis and build a concept catalog. See links submitted.

Project Title: A Practical and Interpretable Approach to Bias Mitigation in Natural Language Generation

PI: Lalana Kagal

Project Summary

Although large language models (LLMs) have demonstrated their effectiveness in a wide range of applications, they have also been observed to perpetuate unwanted biases present in the training data, potentially leading to harm for marginalized communities. In this project, we mitigate bias by leveraging small biased and anti-biased expert models to obtain a debiasing signal that will be added to the LLM output at decoding-time. This approach combines resource efficiency with interpretability and can be optimized for mitigating different types of bias, depending on the target use case. Experiments on mitigating gender, race, and religion biases show a reduction in bias on several local and global bias metrics while preserving language model performance.

Problem Being Addressed

Large language models (LLMs) have been reported to capture and reproduce unwanted biases and stereotypes. This occurs mainly because the large text corpora required for training such models are extracted from the Internet, which is not an accurate reflection of the diversity of real-world distributions. Generating biased outputs can result in serious negative consequences to society, ranging from offensive language that prevent certain demographic groups from adopting the technology to biased job advertisements that
discourage candidates from applying to certain positions. Common approaches to debiasing involve curating better training data or improving the training process. These approaches are time-consuming and resource (both human and computation) intensive. We developed a resource efficient approach that is also interpretable.

Project Status

We developed a framework that mitigates bias by leveraging small expert and anti-expert models to produce a debiasing signal that is then incorporated into the target LLM at decoding-time. This method combines several advantages, including resource efficiency, interpretability, and the ability to customize for specific applications. Throughout the experiments, we observed strong performance-fairness tradeoffs for the framework, out-performing prior research in terms of interpretability and language model performance.

Our system shows promise in generalization since mitigating bias in one direction does not exacerbate bias for others - a property that real-world bias mitigation systems must possess to scale. Investigating the probability shift after debiasing, we provided deeper insights into the

performance-fairness tradeoffs and concluded that results follow expectations. We believe that this framework represents a significant step towards mitigating bias in real-world applications.

We have a paper under review at ACM FAccT conference 2025 (https://facctconference.org/2025/)

Papers Published

A paper under review at https://facctconference.org/2025/

Student Support

A semester or so of one student. The complete project required 2-3 students and the PI's time for around 1-1.5 years.

Source Code

We are waiting for author notification before making the repo public

Project Title: Addressing Issues Related To Trustworthy, Responsible AI, Hallucination-Mitigated Models

PI: Pete Szolovits / Amar Gupta

Project Summary

During the year, we worked primarily on the topic of: Detect and Mitigate Bias, Responsible and Fair AI. To detect bias, we proposed to focus on anomalous results, where the same model produces different predictions for a dataset based on differing sensitive variables. The idea of evaluating the impact of different values of each sensitive feature on the model fairness (and potentially on the model performance) can be studied by leveraging a loss function to measure the impact and develop bias mitigation strategies. We will explore how different types of biases identified with this method can inform the selection of mitigation strategies, considering correlations and data patterns inherent to the datasets.

Moving from a population view on bias to individual considerations, we propose to leverage pre-processing and in-processing techniques to mitigate bias. More specifically, if outlier populations drive the identified bias, we can leverage the situation testing approach to reduce class label bias. Situation testing can be used to test the influence of these combinations of sub-labels on the fairness and, potentially, the performance of a machine learning model.

Further, we propose to leverage multiple approaches to provide fairness guarantees and produce fair ML model outcomes. We plan to address the issue of identifying and mitigating bias and performing experiments on one or more datasets by working closely with one or more MLA member companies.

Problem Being Addressed

The development of new LLM and other AI-based techniques and the growing trend to deploy them in critical financial, technical, medical and other applications have led to a fast growing need to processes and approaches that can address multiple factors that frequently come at the cost of each other. For example, high accuracy of results may come at the cost of comprising the private data provided by customers to the particular bank or other organization. In fact, Customer churn prediction has become crucial for businesses, yet it poses significant challenges regarding privacy preservation and prediction accuracy. (Customer churn refers to the percentage of customers who closed their accounts during a particular time period.)

Further, industries and businesses have a growing need and urgency to seek machine learning models which are more fair and unbiased. This in turn demands careful exploration of several interacting issues, such as how synthetic data and differential privacy affect fairness while maintaining model performance.

Project Title: Addressing Issues Related To Trustworthy, Responsible AI, Hallucination-Mitigated Models (continued from above)

Project Status

We developed and evaluated two approaches to make machine learning models more fair and unbiased:

· A Minimax Pareto-optimized solution, which is a mathematical optimization technique to find the best possible balance between competing fairness and accuracy objectives, while ensuring no single criterion is sacrificed for another;

· A Reject Option Classification (ROC) framework which directly optimizes fairness and accuracy trade-offs without relying on fixed weight adjustments. This is done by introducing a rejection region in the model's decision boundary. When predictions fall within this uncertain region, the model defers decisions to a separate fair classifier, effectively creating a dynamic system that adapts to fairness - accuracy objectives.

Both these methods were tested on real-world, publicly available datasets such as Adult Census Income, COMPAS, and German Credit to evaluate the impact and robustness of our framework. The model's decision-making process improved, leading to an accuracy of 94%.

We also developed FairRAG: a robust architecture that integrates differential privacy, retrieval-augmented generation, and open-source large language models. FairRAG creates synthetic training data using differential privacy techniques, ensuring robust protection of sensitive user information while preserving data utility. At its core, FairRAG utilizes Meta's open-source large language model, OPT-125M, which is augmented with a sentence transformer for semantic similarity matching to retrieve relevant knowledge from a user's profile database. This combination enables FairRAG to enhance prediction accuracy significantly across diverse datasets (bank, gym, and telco customer churn). Our results showed improvements of up to 24% (Bank dataset: from 56% to 80%), 21% (Gym dataset: from 49% to 70%), and 18% (Telco dataset: from 56% to 74%) in prediction accuracy. These gains were maintained when using differentially private synthetic data. Our preliminary results shows FairRAG's ability to build a strong balance between privacy and utility.

We need to find new ways in which one can take a new problem and dataset and quickly determine which of the de-biasing techniques would be most appropriate for this situation. Further, we need to find optimal approaches to integrate two or more of the relevant approaches. So far, we have focused on data from public domain. It would be better to test and adapt these approaches with data from one or more sponsor companies

Papers Published

· “Optimizing Fairness and Accuracy: A Pareto Optimal Approach for Decision-Making” by R. Nagpal, R. Shahsavarifar, V. Goyal, A. Gupta, AI and Ethics, Springer Publishers, July 2024, https://doi.org/10.1007/s43681-024-00508-4

“A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models” by R. Nagpal, A. Khan, M. Borkar, A. Gupta, Machine Learning and Knowledge Extraction, September 2024, 6(3), 2130-2148; https://doi.org/10.3390/make6030105 and https://techxplore.com/news/2024-11-ai-fair-accurate-framework-binary.html binary.html?utm_source=twitter.com&utm_medium=social&utm_campaign=v2

· “FairRAG: A Privacy-Preserving Framework for Fair Financial Decision-Making” by I. Kommula, R. Nagpal, A. Gupta (to be submitted for review during February 2025.)

Student Support

4 students

Source Code

The code will be placed soon on GitHub and the sponsor companies will be alerted about this.

Associated Documents

https://drive.google.com/open?id=1ndlru7rxpMPoK3asYoMBjz7mys-c6ZCl,
https://drive.google.com/open?id=1rssvC-x7rnsfviy3BQfnjl-Zgoo4UQbl,
https://drive.google.com/open?id=1kXVMlk_iwDDIj6kF5kN_VkQghLORwILp,
https://drive.google.com/open?id=1P6Hr-cRHq0-2xLVOnDvikVuqtZRdoLF0,
https://drive.google.com/open?id=1FnUy08KOFd3_SivxeF_bQXkS_d4k55EB

Project Title: Combinatorial Foundation Models

PI: Pulkit Agrawal

Project Summary

With the emergence of foundation models (e.g., LLMs, diffusion models, etc.) that specialize in different facets of intelligence, from visual perception to motor control, we propose a paradigm to dynamically merge specialized models (visual, commonsense, planning, motor, etc.) to address complex tasks requiring diverse expertise. By integrating and reusing knowledge from multiple models, our approach enables the flexibility and robustness needed for long-horizon real-world problem-solving with robotics.

Problem Being Addressed

Many industries need systems (e.g., robotics systems) that can flexibly handle diverse, changing conditions over long time horizons, such as complex assembly lines, logistics, or warehouse operations. Yet current solutions often rely on isolated, single-domain models that excel in that domain but struggle to generalize beyond narrow tasks. By seamlessly integrating specialized foundational models (for vision, language understanding, motor control, and more) into a unified system, we target the core industry challenge of creating robust, adaptive systems that can solve more complex, long-horizon tasks in the real-world.

Project Status

This current project completed, but follow-up works ongoing.

Papers Published

Compositional Foundation Model for Hierarchical Planning (Han*, Ajay*, Du* et al. NeurIPS 2023)

Student Support

Yes, 1