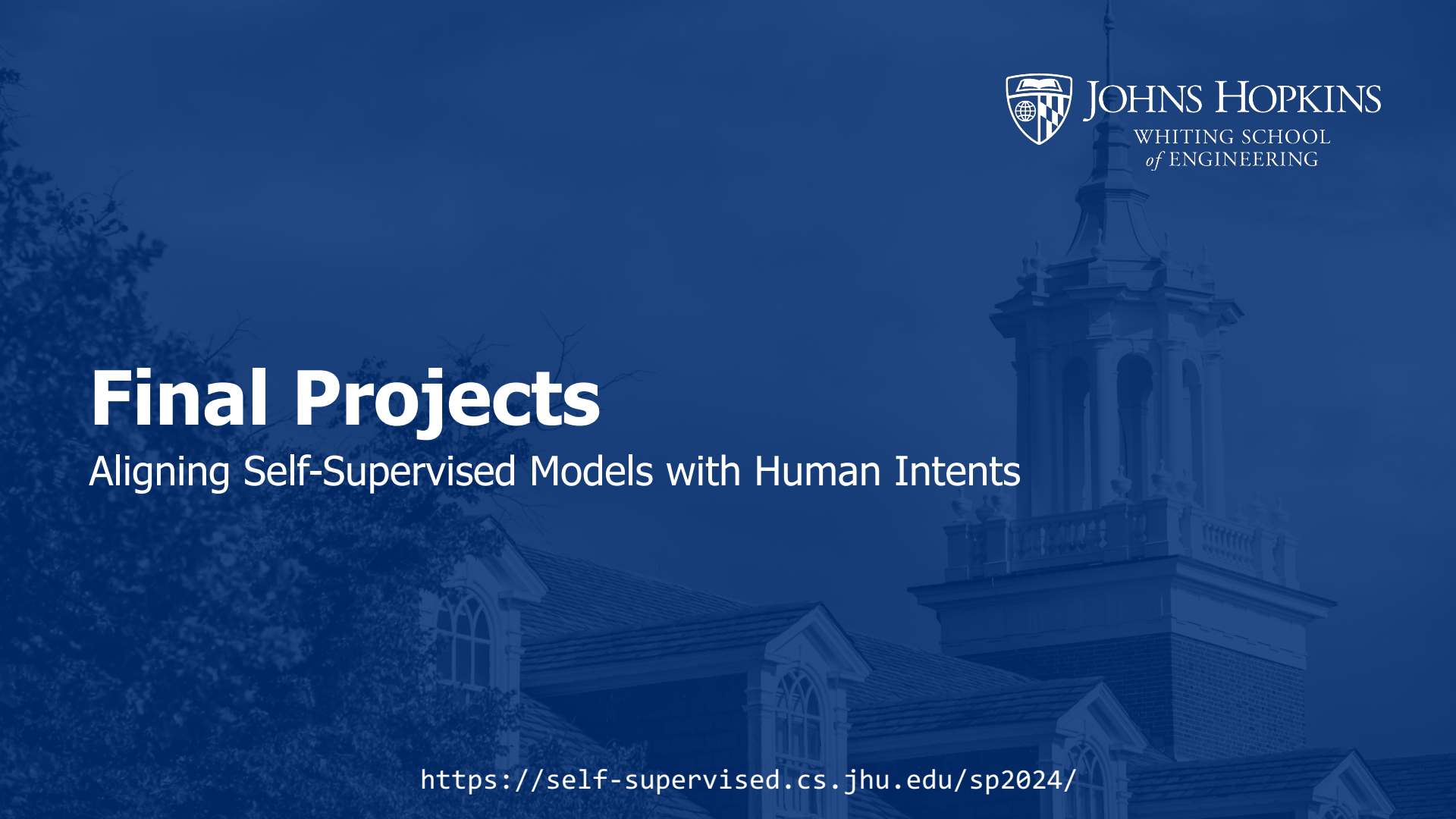


Final Projects

Aligning Self-Supervised Models with Human Intents

https://self-supervised.cs.jhu.edu/sp2024/

Logistics update

- Quiz 2 is this Thursday.
- Will cover everything until the end of last week's class.

- Distribution:
  - 40%: multiple choice
  - 40% short-answer
  - 20% analytical questions

Overview

- 30% of your overall grade (proposal, midway presentation, final report and poster)

- The objective of the final project is to
  - make use of **what you have learned** during this course
  - to solve **a hard problem**.

- What is "a hard problem"?
  - Go beyond what has been done.
  - Identify their weaknesses or make them better. Or make something completely new! (more in a bit.)

- What is the expected amount of effort for the project?
  - Final project effort = ~[4-6] x homework assignment effort

Overview: Topic

- The topic of this project is **open-ended**.

- This project, for example, can focus on
  - demonstrating systemic limitations of prior work or
  - suggesting improvements on methods
  - …

- Really, anything!
- Must substantively involve **human language**!
- We also have a set of **default project ideas**. **Will be released this week.**

Overview: Honor Code

- **Group work:** Students can work in groups (team sizes 1-3 people).
  - Being in a team is encouraged.
  - A larger team project or a project used for multiple classes should be broader and involves exploring more models/tasks/analysis.

- If multi-person: Include a brief statement on the work of each teammate
  - In almost all cases, each team member gets the same score, but we reserve the right to differentiate in egregious cases.

- It's okay (and encouraged) to use existing code/resources
  - Don't re-invent the wheel.
  - You must document it and give them credit .
  - You will be graded on your value-add.
- Use any language/framework for your project. Though we expect most will use PyTorch.

Overview: Mentorship

- You have access to the mentorship of the course staff throughout!

- You will be aligned with at least one (possibly two) course staff.
  - I would love to talk to each team, alas there is only one of me …

- You must meet them **at least once every 2 weeks** to discuss your progress or hurdles.
  - It's your 👊 job to reach out to them to coordinate times, in case office hours do not fit your schedule.

Project Timeline

- Proposal deadline: March 28, via Gradescope
- Midway report: April 16, via Gradescope
- Midway presentation: April 16-25, in class
- Final project posters: May 13 (6-9pm).
- Final project report: EOD, via Gradescope

Project Proposals

- All groups will be required to submit a project proposal.
  - via Gradescope.

- The project proposal is a 2-page description of what you intend to do
  - motivation,
  - hypothesis,
  - experiments,
  - datasets,
  - methods,
  - expected outcome,
  - etc.

Project Proposals

- How to think critically (about an existing paper, application, etc.)
  - What were the main novel contributions or points?
  - Is what makes it work something general and reusable or a special case?
  - Are there flaws or neat details in what they did?

- You need to have an overall sensible idea:
  - Do you have appropriate data or a realistic plant to be able to collect it in a short period of time.
  - Do you have a realistic way to evaluate your work.
  - Do you have appropriate baselines or proposed ablation studies for comparisons.

- Ask for help if needed
  - I encourage you to talk to all course staff about your project ideas.
  - If I had ∞ time, I would talk to all of you. Sadly, that won't happen.

Project Proposals

- Make sure that you follow the expected protocol for the proposal.

  - The project proposal is a 2 page description of what you intend to do
    - motivation,
    - hypothesis,
    - experiments,
    - datasets,
    - methods,
    - expected outcome,
    - etc.

- If you're missing these details, we will not receive the "proposal" credits and we will ask you to redo it.

Proposals: Be as Precise as You Can

"we will first collect an extensive dataset from various sources such as Google News articles, Reddit discussions, and Twitter conversations."

What data? What annotations?

"the model will be trained to produce a sentiment indicator that ranges from -10 to 10 …"

What model?
Where did we get these labels?

You want to be as clear and as specific as possible.

Proposal: An Example

Project Proposal: Ensemble Domain-Specific Knowledge Distillation

Camden Shultz, Kevin Kim, Sara Ren

Motivation:

Even with recent advances in natural language processing and machine learning, large language models (LLMs) tend to suffer from two big problems: they're too big and they hallucinate. Especially as LLMs become more and more widespread in the general population, there will be a greater need for more memory efficient and accurate models, especially those that are capable of answering more specific questions in many different topics. We aim to mitigate the issues of overly large size and hallucinations with one solution: domain-specific knowledge distillation. Knowledge distillation is the process of training a smaller model (the "student") on the input-output pairs of the larger (the "teacher") model. Domain specific distillation requires reducing both the parameter size and vocabulary size while fine tuning to restricted domains of knowledge, such as law, medicine, math, or computer science, which allows models to perform more accurately in domain-specific answering than domain-agnostic distillation. Using knowledge distillation, we hope to construct an ensemble of smaller and more efficient domain-specific and domain-agnostic models that performs similarly to a large model on domain-agnostic tasks, while also reducing hallucinations compared to the larger models for specific tasks.

https://self-supervised.cs.jhu.edu/sp2024/files/proposal-2_seek-extracted_171762894.pdf

Pro

Project Plan

Hypothesis & expected outcome:

We assume that a lot of large language models are currently filled with redundant parameters. By using a combination of domain-specific and domain-agnostic distillation to train smaller and more parameter-efficient language models, we can mitigate the issues associated with large language models, such as their size and tendency to hallucinate, while maintaining performance on domain-agnostic tasks. Recent demonstrations, such as GPT-4 and large language models have shown success in leveraging model capacity to place among higher percentiles in standardized tests. We believe that an ensemble of domain-specific knowledge distilled models can achieve comparable, if not higher, performance on domain-specific questions while maintaining comparable results on domain-agnostic questions and will have an overall lower memory footprint.

Experiments:

1. **Models:** We will use GPT as our teacher model, and if time permits, OPT.
    a. Inspired by our reading, we intend to use an attention-layer over the ensemble for gating queries. Other controllers, including simple classifiers, may also be explored.
2. **Benchmark:** Evaluation of large teacher models on standardized benchmarks and domain-specific knowledge/Q&A.
3. **Distillation Baseline:** Train a single student model instead of an ensemble of student models, evaluating the effects of model compression and distillation.
4. **Expert Student Ensemble:** Evaluation of ensemble of domain-specific student learners on standardized benchmarks and domain-specific Q&A.
5. **Ablation Studies:**
    a. **Global Vocabulary:** Train student models using input-output pairs from the dataset, without restricting vocabulary to a specific domain. This would evaluate the impact of vocabulary trimming on the student model.
    b. **Non-Ensemble:** Train a single student model instead of an ensemble of student models, evaluating the effects of model compression and distillation.

Proposal: An Example

Success Metrics: Evaluate and compare to accuracy of LLMs on standardized tests meant for human test-takers (Olympiad, Bar exam), compare size of ensemble model to LLM.[5]

Datasets:

We think that StackExchange will be an extremely valuable dataset,[6] since it is a huge repository (24 million questions and 35 million answers) of labeled, carefully moderated question/answer pairs that are separated by topic. For domain specific knowledge, we can additionally refer to textbook/wikipedia materials (or other publicly available encyclopedic data) that are available online. The goal is to fit strongly to a specialized corpus (*forum*) for each distilled model, so any dataset with a large amount of truth will work. That is, our transfer set will be task-specific, rather than task-agnostic.

Halfway Milestone:

In our halfway milestone, we hope to have a pipeline for a single downstream task complete that allows us to have a specialized distilled model on Math Olympiad questions. By having this pipeline complete, we can then generalize to other tasks with relative ease (where the only limiting factor will be training time) and focus on evaluation as well as ablation.

Timeline: Midway Report

- This is a progress report — Max 5 pages, via Gradescope.

- Should be more than halfway done!
- Describe the progress made
  - Experiments you have run
  - Preliminary results you have obtained
  - how you plan to spend the rest of your time.
  - ….

- You're expected to **have implemented *some system***, and to have ***some* experimental results** to show by this date.

Final Report

- You will write up the results in a report.
  - Should be max 8 pages, on Gradescope.
- The final report should summarize your findings and answer the following questions:
  1. **Motivation:** What approach did you take to address this problem, and why?
  2. **Related work:** How did you explore the space of solutions?
  3. **Approach:** How did you pick your solution?
  4. **Evaluation:** How did you evaluate the performance of the approach(es) you investigated?
  5. **Findings:** What worked, what did not work, and why?
  6. **Conclusion** and future work

- Writeup quality is very important for your grade!
- See course page for an example of a strong report.

Final Poster

- All students will present their findings at a poster presentation during the final exam period.

- Then we will celebrate! 🎉

- Best Project Awards !
  - Selected by course staff
  - Popular vote by the class
  - Awards details are TBD

Back to finding project ideas

Suggesting a small extension to a paper

- Choose a paper, propose a small **improvement** or extension to the paper.
- The paper should be high-quality
- Ideally discuss your plan with TAs early on selecting the paper In your proposal justify why you chose this paper

Finding Good Papers

- Check out papers from recent conferences:
  - ICLR 2024 submissions (will be available on OpenReview end of September)
  - NeurIPS 2023: https://openreview.net/group?id=NeurIPS.cc/2023/Conference
  - ICML 2023: https://icml.cc/virtual/2023/papers.html?filter=titles
  - ACL 2023: https://aclanthology.org/events/acl-2023/#2023acl-long
  - ArXiv and Twitter (be aware of noise and low-quality papers)

Behavior analysis of models

Analyze the behavior of a model to provide some new insights.

- Examples:
  - Investigate hypotheses like “LLMs are more likely to hallucinate content if they are presented with unfamiliar examples”.
  - Investigations into the data. How training data can impact some behavior of LLMs.
  - Investigating the bias and trustworthiness of LLMs

Theoretical Analysis

- Theoretical analysis: Show some interesting and new theoretical analysis/results of an existing approach.
- This type of project is only recommended for folks who have a strong theoretical background E.g., prove that sparse attention models are universal function approximators.

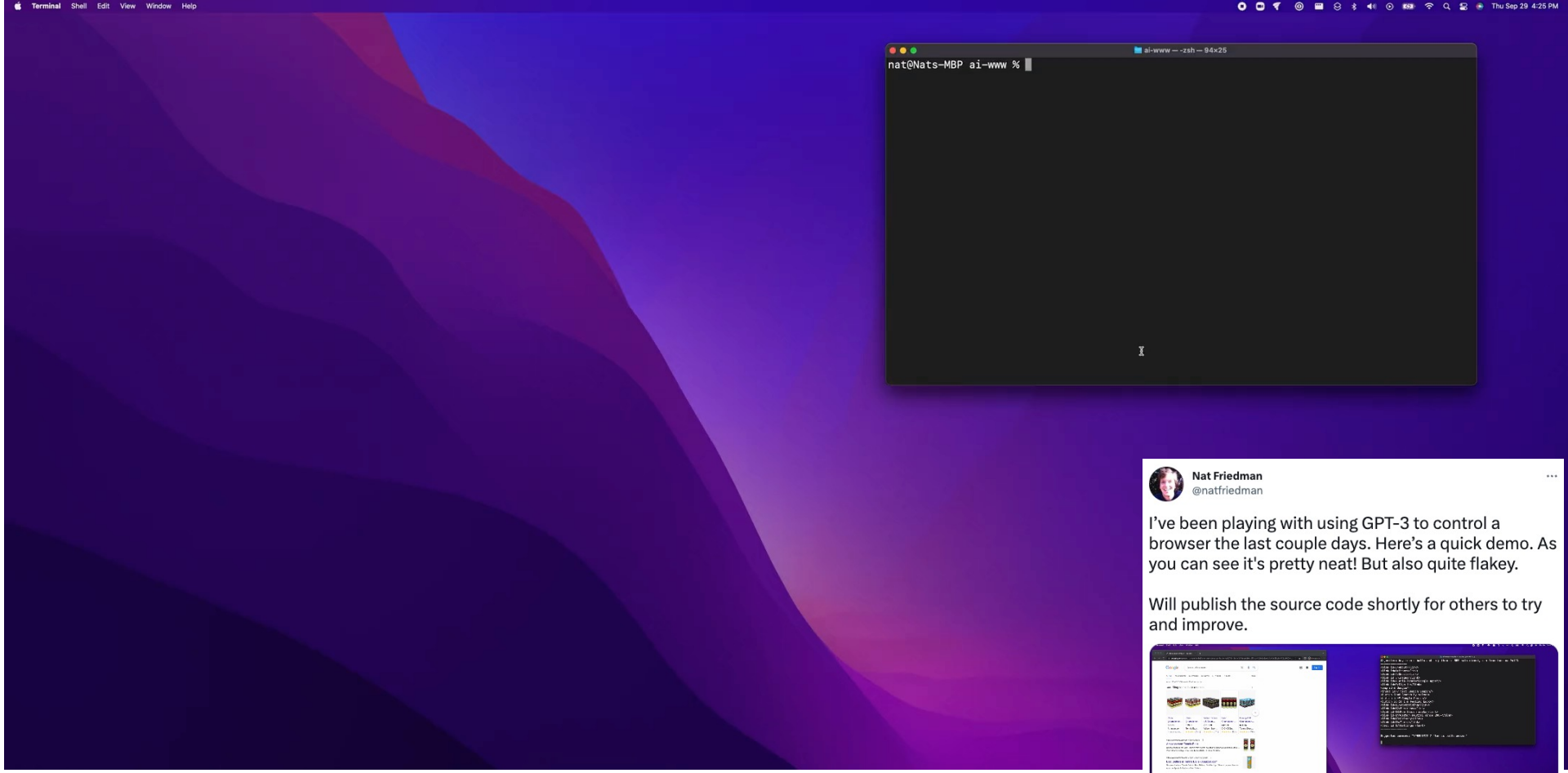
Think Broadly

- Take into account the topics that we will discuss in coming weeks: RLHF, text2code, tool-use, LLMs for grounded reasoning, vision-language models, efficient computational frameworks, bias and fairness, legal issues regarding LLMs.

- Look at an interesting interdisciplinary problem,
- This is a perfect project for those who have other backgrounds in addition to CS
  - E.g., LLMs for scientific discovery
  - E.g., Applications to education
  - E.g., LLMs for Robotics

Applications

- Good project for those who are excited about building applications.
- Here are a few examples:



Nat Friedman
@natfriedman

I've been playing with using GPT-3 to control a browser the last couple days. Here's a quick demo. As you can see it's pretty neat! But also quite flakey.

Will publish the source code shortly for others to try and improve.

7:38 PM · Sep 29, 2022

https://twitter.com/natfriedman/status/1575631194032549888

You're in charge ….

- Lastly, this is your chance to get your hands dirty and work on a problem that you care about.

- Think broadly!