DS-GA 1003: Final Project Guidelines

Project Overview

The final project for this course is a research paper on a topic of your choice. You will form groups of 2-3 students for this project, and may choose either of the following tracks:

- *Applications Track.* Choose a real-world problem, identify how and why machine learning could be helpful, and find or collect a relevant dataset for the problem. Write a survey describing previous approaches to this problem. Then, establish reasonable baselines and compare the performance of several different ML techniques learned in class. Analyze the failure modes of these methods, explain why some methods work better than others, and characterize any remaining gap in task performance.
- *Research Track.* Identify an important gap in the literature for an ML topic of interest, and then propose and execute experiments to address the gap. Write a paper in the style of a top-tier machine learning venue (e.g., NeurIPS, ICLR, ICML, ACL, CVPR). Students are welcome to continue on their pre-existing research projects if they choose to pursue this track.

The latter track is more difficult and is intended for students who are looking to conduct cutting-edge research that could eventually be presented at a top-tier machine learning venue. However, grading rubrics for the two tracks will identical.

In either case, your final paper should be up to **8 pages long**. Your paper may be shorter than this, but you may lose points if you do not satisfy all components of the grading rubric.

Due Dates

Before submitting the final project, you should form groups of 2-3 students and submit a short 2-page proposal outlining your plans for the final project. Your proposal should provide context for the problem, describe your plans for tackling it, and discuss any potential hurdles.

- Groups formed: February 28th, 2026
- Proposal submitted: March 31st, 2026
- Paper submitted: May 8th, 2026

Each component will be due at 11:59PM ET. See below for submission details.

Rubric

We anticipate that most final projects will be empirical in nature. If you are conducting an empirical final project, it will be graded as follows:

- *Proposal (10%).* Does the proposal describe a clear problem and approach?
- *Problem selection (10%).* Does the project involve a novel, interesting, and/or impactful problem of interest? Is the problem domain clearly described and well-motivated?
- *Related work (20%).* Does the paper appropriately survey related work, including both existing datasets and methods? Does the paper clearly and accurately describe how the current work is similar to or different from prior work?
- *Data (10%).* Does the paper clearly describe the data used in the project? Are basic data statistics (e.g., number of samples, distribution of labels) provided?
- *Methods (10%).* Does the paper clearly describe the method(s) used? Are the methods clear, straightforward, and well-motivated?
- *Results (20%).* Are the results clearly presented with tables and/or figures? Are high-level takeaways clearly summarized? (Negative results are fine.)
- *Analysis (20%).* Does the paper provide an in-depth analysis of results? For example, in a classification task, does the paper describe which datapoints models misclassify?

If your final project is not empirical (e.g., in learning theory), then it will be graded on correctness, clarity, and novelty. We expect that most papers will be empirical in nature, but feel free to reach out to the instructors for more info on a more theoretical paper. If this is of interest to you, some top-tier conferences that focus exclusively on theory (as related to machine learning) are COLT and ALT, although many theory papers also appear at NeurIPS and ICML.

Note: We highly encourage (but do not require!) you to checkin with the instructors and TAs regarding your final proposal idea, your planned experiments, and your writeup. We anticipate that everyone who regularly checks in with the course staff and follows their recommendations should be able to receive an **A** on the final project.

Submission

Submission for group formation, proposals, and the final paper will be available on Gradescope. Please only make **one submission per group** for each of these three components.

How to Choose a Topic

We've received many questions about how to pick a topic for the final project. If you're pursuing the applications track, then we'd recommend first trying to identify a potential domain of interest (e.g., predicting ridership on the MTA), then try to find whether any data for this problem already exists or could be collected with a reasonably small amount of effort. Just to get you started, here are some datasets that are available publicly on the internet:

- UC Irvine Machine Learning Dataset Repo.
- NYC Open Data.
- Data.gov.
- Kaggle.
- Newsgroup dataset.
- AWS Open Data Repo.

If you're pursuing the research track and haven't done ML research before, then here are a few suggestions:

- Check out conference workshops! Conferences like NeurIPS, ICML, and ICLR typically have 20+ workshops spanning a range of topics. Scrolling through these topics to see which ones interest you can be a good way to find a general topic area of interest, and individual workshops will often have papers or invited talks which might give you a sense of recent research in that subarea.
  - Workshops from the most recent (2025) ICML.
  - Workshops from the most recent (2025) NeurIPS.
  - Workshops from the most recent (2025) ICLR.
- Read technical blogposts! Blogs can provide a good big-picture introduction to a topic of interest. For example, Lilian Weng and Jacob Steinhardt both write high-quality technical blogposts. Labs like BAIR also often have high-quality technical blogposts. As a general guideline, blogposts on websites like Toward Data Science and Medium tend to be of lower quality, although there are plenty of exceptions.
- Come chat with course staff! The instructors and TAs are actively conducting ML research across a range of topics, and we'd be happy to tell you about our areas of interest. One useful exercise could be to bring a list of potential research topics to office hours and discuss them with staff members. For more guidance, see here: `https://colah.github.io/notes/taste/`

Compute

For students in the applications track, websites like Google Colab and Kaggle Notebooks often provide sufficient compute resources. For final projects which require more compute, however, we have obtained a small amount of HPC Cloud Bursting credits through NYU.

Cloud Bursting access is provided through Open OnDemand (OOD). This allows you to:

- Launch compute nodes
- Run Jupyter notebooks
- Open terminal sessions
- Transfer some data from local computers
- Submit batch jobs directly from Jupyter notebook terminals

Note that the NYU VPN is required to access these nodes when working off campus.

Slurm Account and Resource Allocation Students must use their assigned Slurm account for this course. Each student is assigned the following Slurm account:

- **Account:** `ds_ga_1003-2026sp`
- **GPU allocation:** 300 GPU hours (18,000 minutes)
- **CPU allocation:** Sufficient CPU time for coursework

Allowed partitions:

- `interactive`
- `n2c48m24` → CPU only
- `g2-standard-12` → 1 L4 GPU
- `g2-standard-24` → 2 L4 GPUs
- `g2-standard-48` → 4 L4 GPUs
- `c12m85-a100-1` → 1 A100 40GB GPU
- `c24m170-a100-2` → 2 A100 40GB GPUs
- `n1s8-t4-1` → 1 T4 GPU

Spot Instance Policy Cloud resources are running on Google Cloud spot instances, which may be preempted at any time. More information can be found here:

- https://cloud.google.com/compute/docs/instances/spot
- https://cloud.google.com/compute/docs/instances/preemptible

Best practices:

- Enable checkpoint/restart for production runs
- Save checkpoints to your `/scratch/$NETID` directory
- Jobs will be automatically requeued if instances are shut down by GCP

Add the following directive to all Slurm scripts:

```
#SBATCH --requeue
```

Policy on the Use of LLMs

Overreliance on LLMs can have negative impacts on learning and skill formation (Shen and Tamkin, 2026). As a result, the use of LLMs to write final projects is forbidden.

However, tools like Cursor and Claude Code are proving increasingly useful in modern ML research workflows. Therefore, *for the research track only*, you are allowed to use AI coding assistant tools. However, you are responsible for verifying the correctness of AI-generated code. Regardless of which track you choose, you may not use LLMs to write the paper itself.

References

Shen, J. H. and Tamkin, A. (2026). How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*.