👋 Intro

This tab contains all onboarding info. Please go through it thoroughly.

TODO (all team members)

Please fill this out as you complete it.

Name	Diogo	Alexander	Ashwin	Jan	Shariqah	Yeonwoo
Join Slack channel	☑	☑	☑	☑	☑	☑
Check you have access to Google Drive folder	☑	☑	☑	☑	☑	☑
Check you have access to Gathertown	☑	☑	☑	☑	☑	☑
Add Github username below	☑	☑	☑	☑	☑	☑
Check you have access to Google calendar	☑	☑	☑	☑	☑	☑
Read Intro tab	☑	☑	☑	☑	☑	☑
Check your bio	☑	☑	☑	☑	☑	☑
Fill out both when2meet (Preferred, Extended)	☑	☑	☑	☑	☑	☑
Indicate group preferences	☑	☑	☑	☐	☑	☑
Brainstorm what you expect to get out of the project and possible failure modes	☑	☐	☐	☐	☑	☑

- Join Slack channel: you should have been invited to a private group channel by the organizers, called **diogo-cruz-eval-framework-for-llm-unlearning**. If not, check with Diogo.
- Check you have email access to the Google Drive folder SPAR Unlearning Project.
- Check you have access to Gathertown: we'll use it for coworking sessions.
- Add Google calendar: Diogo will add the weekly meetings and coworking sessions to it. You are also free to use it to schedule any other useful meetings between team members.
- Read Intro tab (like you're doing now 🙂)
- Check your bio, correct any mistakes there.
- Fill out both when2meet: the preferred option is for times where you're generally available, the extended option is for times that you could make work for meetings if necessary, but you'd prefer to avoid. **Note that the times on when2meet are in UTC,**

not your timezone. If you have changing availability, then focus on your expected availability for the next few weeks. **I'll also use your stated availability to schedule the kickoff meeting during the first week (Feb 10 to 15)**.

- Indicate group preferences below.
- Brainstorm what you expect to get out of the project and possible failure modes in the **Expectations and Failure modes** tab.

Team

We'll also get to know each other better during the opening weekend.

Name	UTC	Bio	Github (email)
Diogo Cruz	0	Has a technical background in physics (PhD in quantum computing) and recent experience in AI safety research, currently working on Evals related to unlearning, multi-turn jailbreaks, and autonomous agents.	diogo-cruz (diogo.abc.cruz@gmail.com)
Alexander Panfilov	+1	PhD student at IMPRS-IS/ELLIS focused on adversarial robustness and jailbreaking attacks against LLMs, with publications at major ML conferences including ICLR 2024 and NeurIPS 2024 Workshop.	Kotekjedi (kotekjedi@gmail.com)
Ashwin Sreevatsa	-8	Software engineer at Google specializing in ML workflows and model lineage, with research publications in NLP and medical imaging, and active engagement in machine unlearning research through systematic literature review.	TheAshwinner (public.ashwinsreevatsa@gmail.com)
Jan Batzner	+1	PhD candidate at TU Munich investigating LLM auditing methodologies, with prior experience as SPAR mentor and active contributions to AI safety initiatives including BASIS and AI Safety Camp.	Janbatzner (jan.batzner@gmail.com)
Shariqah Hossain	-5	Recently completed M.Eng. in EECS with specialization in AI where she worked as a MIT CSAIL researcher evaluating model editing algorithms for knowledge unlearning in LLMs and has a background in healthcare data system engineering.	shariqahn (shariqah97@gmail.com)
Yeonwoo Jang	-5	Senior Research Data Scientist at Bank of America with expertise in AI interpretability, former AI researcher in medical imaging; holding degrees from the University of Oxford (MSc Statistics & Machine Learning) and Imperial College London (BSc Mathematics).	yoenoo (yeonwoojangus@gmail.com)

Project structure

Here's the general structure of the project:

- SPAR Unlearning Project is the folder that contains all the files relevant to the project.
    - Exceptions are the Github repo(s), and, later on, the paper draft files (we'll likely write the final draft in Overleaf and potentially use Zotero for the references).
    - By default, when you add a file to the folder, it should be shared with everyone else. You can change the sharing permissions to remove others, if you want a personal file (for notes, for example).
- This file (SPAR Project notes) is the main one for the project. Anything that is expected to be useful for everyone else should go here. You can create separate tabs for different topics. Currently, it contains the tabs:
    - **Intro**: this one.
    - **Meeting notes**: notes for the weekly team meetings we'll have, and any coworking and asynchronous notes. In general, **treat this tab as a Logbook for the project**.
    - **Resources**: useful links for the project.
    - **Original proposal**: copy of the original proposal, for easy reference, and in case we want to modify it.
    - **Literature review**: collection of links that might be pertinent to the project.
    - **Failure modes**: list of failure modes for the project, and ways to address them.
    - We can also create other tabs for different topics, for coworking sessions, etc. Feel free to modify this doc as you wish.
- You can create separate docs for things that are only relevant for some team members.

Initial Setup

We'll have a 3h kick-off meeting the first week, if possible.

For the first few weeks (1–3), we'll divide ourselves into **2–3 focus areas**. Each group will have concrete tasks and deliverables, making it clear how their work fits into the overall project:

Group L: Literature & Baselines

Goal: Ensure we have a solid understanding of existing unlearning methods, their code (if available), and their current evaluation approaches.

- **Tasks**
    1. **Deep Literature Review**
        - Identify and summarize relevant unlearning papers (e.g., LLMU, RMU, distillation-based unlearning, entity-level unlearning).

        - Capture the main ideas, algorithms, limitations, *and any code bases* that might be publicly available.
    2. **Method & Metric Inventory**
        - List out the existing evaluation metrics these papers use (e.g., forget-set performance, utility preservation metrics).
        - Note down any open-source tools or scripts that can be readily leveraged.
    3. **Reproduce a Simple Baseline**
        - Select at least one known unlearning approach (preferably one with a usable open-source implementation) and try to replicate its results on a **small** toy example.
        - Document steps to run it, any dependencies, and summarize how well it works "out of the box."
- **Deliverables**
    1. Concise "Lit Review & Methods Matrix" (what methods exist, how they differ)
    2. A brief internal "Baseline Reproduction Report" (less than 2 pages or a Google Doc summary) describing results, pitfalls, and suggestions for improvement

Group B: Benchmark Creation & Data Curation

Goal: Gather and structure the test datasets (or create new ones) that we will use for evaluating unlearning techniques in LLMs.

- **Tasks**
    1. **Identify/Collect Existing Benchmarks**
        - E.g., WMDP for harmful knowledge, TOFU for privacy data, or other relevant sets.
        - Pull them into a single place (e.g., a shared folder or a GitHub repo).
    2. **Clean/Format Benchmarks**
        - Standardize dataset formats so they're easy to plug into our future pipeline: e.g., JSON lines with `(prompt, expected_output, context)` or multiple-choice format.
    3. **Design Simple "Starter" Test Cases**
        - Draft a handful of small, self-contained examples—for instance, a mini version of "how to build a harmful device" or "private data about John Doe"—so that we can quickly test any prototype unlearning method.
    4. **Optional**: Start brainstorming **synthetic** expansions or new prompts that systematically probe unlearning (e.g., paraphrased queries, multi-step prompts).
- **Deliverables**
    1. **Organized Datasets** (in the shared Google Drive or GitHub) with a README describing how to load them
    2. A short "Benchmark Inventory" doc summarizing the coverage (e.g., "WMDP: 3,668 QA items on illicit instructions," "TOFU: synthetic personal data QAs," etc.)
    3. A small "Starter Cases" folder with examples for quick testing

Group M: Metrics & Pipeline Development

Goal: Begin implementing an actual testing/evaluation pipeline—collecting code tools, writing scripts, and drafting how we'll measure unlearning success.

- **Tasks**
    1. **Metrics Specification**
        - Work with the Literature & Baselines Group to refine which metrics we'll implement first (e.g., simple "accuracy drop" on the forget set, "retain-set performance," plus something for adversarial prompts or robustness).
    2. **Skeleton Code for Evaluation**
        - Set up a basic Python package or Jupyter notebooks that can load a model, run prompts from a dataset, and compute the chosen metrics.
        - Possibly create a single script (`eval.py`) that can be extended later.
    3. **Preliminary Dashboard/Visualization** (Optional)
        - If time allows, start drafting a minimal "dashboard" (e.g., a Jupyter notebook or Streamlit app) that visualizes forgetting vs. retention.
    4. **Collect Infrastructure Requirements**
        - Decide how we'll run the pipeline (local GPU vs. cloud).
        - Clarify if we need to containerize anything (e.g., Docker) or set up a GitHub Actions workflow for automated testing.
- **Deliverables**
    1. A **basic script** or notebook that can:
        - Load a model (even a small one)
        - Run it on a few "forget" and "retain" prompts
        - Print out or log relevant metrics (accuracy, perplexity, or whichever is easiest to do first)
    2. A short "Metric & Pipeline Doc" explaining the evaluation flow, how to add new metrics, etc.

Coordinating Across Groups

- **Shared Weekly Meetings and Coworking sessions**: Each group reports on progress and blockers.
- **Cross-Group Checkpoints**:
    - **Literature & Baselines** feeds their results/insights to **Metrics & Pipeline** so we know which metrics matter most to replicate.
    - **Benchmark** group provides sample data to **Metrics & Pipeline** group to test code.

In Diogo's experience, we'll probably quickly deviate from the initial setup, but it's a good starting point to ground the project.

Please mark your preferences in this table (from 5 for "I'd love to work on this", to 1 for "I'd prefer not to work on this", with 3 being neutral). This is non-committal, it's just to get a sense of

everyone's preferences. We'll sort things out during the kick-off meeting. In practice, you are free to move from group to group as we progress through the project.

Name	Group L	Group B	Group M
Diogo (L)	5	2	2
Alexander (M)	3	3	4
Ashwin (B/M)	3	5	4
Jan (L)	5	5	3
Shariqah (B)	3	5	4
Yeonwoo (M)	3	5	5

Communication

- Be respectful and kind in all interactions, acknowledging that team members have different working styles and preferences.
- Maintain active communication - share your ideas openly and encourage input from others, especially those who may be quieter in discussions.
- Stay responsive on Slack and attend scheduled calls. If you can't meet commitments, communicate this proactively to allow the team to adjust.
- If interpersonal issues arise, reach out to Diogo or to the SPAR organizers who can help address concerns appropriately.

Team meetings

The default way to communicate is our team channel on Slack. You should also make ample use of the Meeting notes tab to write summaries or updates that are useful to the rest of team

- Please avoid using email unless as a last option.

For recurring weekly meetings: for now, we'll use my personal meeting room: https://mit.zoom.us/j/6119047956. We'll aim for this to be a 1h meeting, but if that proves to be insufficient, we might change it to 1.5h.

For coworking sessions: we'll use Gathertown: https://app.gather.town/app/btWwxHpCKiQ8vuO3/SPAR_unlearning

For 1-on-1's: you can just use the Google Meet link associated with the calendar invite, or Gathertown. When setting 1-on-1's, or impromptu meetings, it's recommended that you mark

them on the Google calendar, and make sure the event is visible to any relevant team members. For Diogo, you can also just quickly schedule a meeting at https://calendar.app.google/yNoZ1yUmQoQ39TxR7.

Timeline

Overview

The project spans from February 10th to May 17th, 2025 (approximately 14 weeks).

Key Dates

- **Feb 7th**: Virtual kickoff for mentees
- **Feb 10th**: Projects officially begin
- **April 3rd**: Midterm report due
- **May 8th**: Virtual poster session
- **May 17th**: Final report due

Month 1 (February 10th - March 10th)

Focus: Project setup, background research, initial metric design, and basic benchmark creation

Month 2 (March 11th - April 10th)

Focus: Framework implementation, pipeline development, and preliminary testing

Month 3 (April 11th - May 17th)

Focus: Full evaluation study, analysis completion, and documentation

Checkpoints and Milestones

Week 5 MVP Checkpoint

Expected deliverables:

- Working prototype of evaluation framework
- Initial set of implemented metrics
- Basic benchmark datasets
- Internal presentation of progress
- Clear roadmap for remaining work

Week 10 External Draft Checkpoint

Expected deliverables:

- Complete evaluation framework
- Full set of benchmark results
- Draft paper for external feedback
- Comprehensive documentation
- Initial findings and analysis

Week 14 Final Deliverables

Expected outputs:

- Research paper draft
- Open-source evaluation framework
- Benchmark datasets
- Documentation and usage guidelines
- Final project report
- Blog post summarizing findings

Meeting Schedule

- Kick-off meeting in week 1
- Weekly team meetings: [TBD based on when2meet results]
- Regular (optional) coworking sessions in Gathertown
- Additional group-specific meetings as needed

Tools

- Compute: SPAR should cover us for up to 500$ in compute.
- My suggestion is that we rent an instance on vast.ai and use it throughout, but we can check other options. The main annoyance with this option is that we might need to reinstall everything whenever we stop the instance.

📝 Meeting notes

Aug 1, 2025

Based on the reviewer feedback and typical camera-ready requirements, here's a comprehensive TODO list for preparing your camera-ready version:

Camera-Ready TODOs

1. **De-anonymization & Formatting**

- [x] Remove "Anonymous Author(s)" and add actual author names
- [x] Add proper author affiliations and contact emails
- [ ] Update the header to reflect COLM 2025 Workshop SoLaR acceptance
- [ ] Remove the "Do not distribute" footer
- [ ] Apply the final camera-ready LaTeX template from COLM 2025

2. **Address Reviewer Concerns**

High Priority (Directly Addressable):

- [ ] ~~Add systematic evaluation of other non-Latin scripts (Bengali, Arabic already tested but emphasize the comparison)~~
- [ ] Provide more rigorous justification for using tinyMMLU (100 samples) or add evaluation on full MMLU
- [x] Clarify the speculation about Hindi tokenization patterns - either provide evidence or soften the claim
- [ ] Add a limitations section discussing what mechanistic understanding is still missing

Medium Priority (Partial Address):

- [x] Add brief discussion of societal/policy implications in Impact Statement
- [ ] Acknowledge the limitation of not having mechanistic analysis of unlearning methods
- [ ] Consider adding references to related work on tokenization effects in multilingual models

3. **Technical Improvements**

- [x] Ensure all hyperlinks work (currently [anonymized for double-blind review])
- [ ] Add the actual GitHub repository URL for the evaluation framework - check this URL too
- [x] Verify all model checkpoint names and links are correct
- [ ] Double-check all numerical results in tables match the figures

4. **Content Polish**

- [x] Proofread for any typos or grammatical errors
- [x] Ensure consistent terminology throughout (e.g., "Hindi filler" vs "Hindi filler text")
- [ ] Check that all figures are high-resolution and readable
- [ ] Verify all citations are complete and properly formatted

5. **Acknowledgments & Funding**

- [x] Add acknowledgments section
- [x] Include any funding sources
- [ ] Thank reviewers for their feedback

6. **ArXiv Specific**

- [ ] Update the arXiv version to match camera-ready
- [ ] Add "Accepted at COLM 2025 Workshop SoLaR" to arXiv comments
- [ ] Ensure the GitHub repository is public before arXiv update

7. **Optional Enhancements (if page limit allows)**

- [ ] Add an appendix with additional language comparisons
- [ ] Include a brief discussion on future work regarding mechanistic analysis
- [ ] Add a summary table comparing all unlearning methods tested

8. **Final Checks**

- [ ] Verify compliance with page limits
- [ ] Run spell checker and grammar checker
- [ ] Have co-authors review the final version
- [ ] Check that all supplementary materials are included

Remember to prioritize addressing the reviewers' main concerns while maintaining the paper's core contributions. The reviewers were generally positive (ratings 7-8), so focus on clarifications rather than major restructuring.

May 14, 2025

Week 14

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	•
Ashwin	☑ ~~Identify the discrepancy between the different probe accuracies (on a first glance, I suspect my 40% number is less trustworthy here)~~ ☑ ~~Poster/presentation?~~	☐ Probe results for ELM as well ☐ Probe for the different prompts	• Editing poster/presentation • Updating diagrams from poster for the draft (still need to upload)
Shariqah	☑ ~~poster/presentation~~	☐ Help Ashwin with probing	• Updated multi shot, model comparison figures • Updated draft with midterm report info • Created script for extracting answered/non-answered prompts • Reached out to authors for RMU weights and more info on best robustness test calculation • Ran preliminary results on ELM probing, but probe still needs tuning
Yeonwoo	☐ Poster/presentation ☐ Run eval w/ LLMU model ☐ Many-shot MMLU ☐ Many-shot bio-retain & bio-forget	☐ addChar analysis 5% vs full set ☐ Probing ☐ Activations + PCA analysis ☐ How sensitive lm-eval results	☑ ~~Run eval w/ LLMU model~~ ☑ ~~Many-shot MMLU~~ ☑ ~~Many-shot bio-retain & bio-forget~~

		on answer choices (e.g. abcd vs (a)(b)(c)(d), etc.)	

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    -

Other (lower priority) Questions:

- Diogo:

- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:

Pre-meet notes:

- Diogo:

Meeting notes

Paper Progress

- Shariqah updated the model comparison figure with a single plot using a colorblind-friendly palette
- Current issues with figures include font sizes being too small (need to be closer to text size)
- Figure 2 may be too small as currently designed
- The team is having trouble replicating results from a paper they're referencing
    - Yeonwoo found significantly different coverage percentages (6% vs 25% claimed in the reference paper)
    - Potential causes: different translation models (Haiku vs Opus), unreported methodology changes
    - The reference paper authors reported they no longer have access to their rephrasing datasets
- The team discussed how to handle this discrepancy in their paper, deciding to use their own results as the baseline

Technical Discussions

- LM model output frequently produces empty strings or incorrect formats when tested
- The team debated whether to pursue probing analysis but decided to skip it unless significant updates appear by Friday

- Discussed examining logit differences between correctly and incorrectly answered prompts
- Considered different threat models: API access (token-only) vs direct logit access

Project Timeline

- This was noted as the final weekly meeting for the project
- A three-hour co-working session is scheduled for Friday
- Final draft and report due next week (Monday)

Next Steps

- Focus on adding content rather than cutting down at this stage
- Two critical missing sections need to be added: setup section and related work section
- Fix references which may contain errors
- Continue improving figures to ensure readability

TODO List

For Yeonwoo

- ☐ Run Bengali translation test cases using Opus instead of Haiku to check for differences
- ☐ Complete logit analysis comparing correctly vs incorrectly answered prompts
- ☐ Update team on findings about the reference paper's results

For Shariqah

- ☑ Make font sizes bigger in figures
- ☑ Fix the multi-shot figure formatting
- ☑ Consider reducing information density in figures (e.g., consolidate rephrasing point types)
- ☑ Draft the related work section
- ☑ Support Yeonwoo with code assistance if needed

For Ashwin

- ☐ Continue working on diagrams for the draft
- ☐ Share works-in-progress regularly in Slack for feedback
- ☐ Help with the setup section
- ☐ Assist with draft conversion as needed

For All Team Members

- [ ] Attend Friday's three-hour co-working session
- [ ] Add content to draft where appropriate (create new appendices if unsure where content belongs)
- [ ] Update the "next day goals" table with current tasks
- [ ] Prepare for final report generation on Friday
- [ ] Consider providing project feedback after Monday

For Diogo

- [ ] Add meeting notes to the project documentation
- [ ] Think about improvements for Figure 1
- [ ] Oversee the completion of missing sections
- [ ] Review current draft progress

Next Day Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☑ ~~Add diagrams/setup section~~	☐ Related work section? ☐ Any other sections that are still todo?	
Shariqah	☐ Update figures ☐ Write draft	☐	
Yeonwoo	☐	☐	

May 12, 2025

Coworking

Final meeting notes

The team discussed preparing their paper for submission to the "MUGen workshop," focusing on adapting it to meet the requirements of the ICML 2025 format (4 pages excluding references

and appendices, two-column format). The paper has already been converted from the original NeurIPS format to the required ICML format.

The group discussed their paper's content and conclusions about unlearning techniques. Their research shows negative results regarding knowledge retrieval through prompting, which contrasts with previous work. They specifically highlighted the distinction between formatting issues versus actual unlearning, comparing their findings to the "learning truly unlearn" paper. They noted discrepancies between their results and previous work, potentially due to different model implementations, different prompts, or varying degrees of unlearning.

The team (Diogo, Yeonwoo, Shariqah, and Ashwin) debated additional tasks like implementing the LLMU model, running probing approaches, and conducting a more detailed analysis of logits versus formatted outputs. They decided to prioritize getting the draft to 80-90% completion before dedicating resources to additional analyses. They also discussed creating a setup section with diagrams to clarify their methodology for readers.

There was concern about the interpretation of LLM eval metrics, which needs investigation to ensure accuracy in their reporting. The team also planned to add a related work section covering 10-15 papers in 2-3 paragraphs.

TODOs

1. **Continue adapting content to ICML format**
    - Ensure content fits within the four-page limit (excluding references and appendices)
    - Review how the content looks in the new two-column layout
2. **Restructure figures** (Shariqah to lead)
    - Adapt images to work well in two-column format
    - Consider condensing multiple figures into a single plot if possible
    - Focus on plots that highlight the main message and move others to appendix
3. **Create setup section with diagrams** (Assigned to Ashwin)
    - Create flowcharts/diagrams explaining the methodology
    - Include examples of prompts, model responses, and logit extraction
    - Build on existing diagrams from the poster presentation
4. **Reorganize content**
    - Move lists and large tables to the appendix
    - Convert tables into plots where appropriate for better legibility
    - Update the draft with the latest information
5. **Additional research tasks**

    - Work on the LLMU model implementation (assigned to Yeonwoo)
    - Investigate how LLM eval calculates metrics (assigned to Ashwin)
    - Analyze logits for correctly formatted vs. incorrectly formatted answers
    - Consider probing approaches after draft is nearly complete
6. **Add related work section**
    - Write 2-3 paragraphs covering 10-15 papers
    - Focus on papers they've followed closely and cited throughout
    - Can be completed toward the end as it's independent of other sections
7. **Address result discrepancies**
    - Add discussion (likely in appendix) about why their results differ from previous work
    - Explain potential factors: different rephrasing, models, or unlearning effectiveness
8. **Timeline**
    - Team members have 2-4 hours each before Wednesday meeting
    - Aim to have draft 80-90% complete before pursuing additional analyses
    - Reconvene on Wednesday to assess progress

Ongoing meeting notes

- Address discrepancy between our results and “truly unlearn” paper

Current Paper Status and Requirements

- Paper needs to be converted from Europs format to ICML format
- Must fit within 4 pages (excluding references and appendices)
- Will require conversion to two-column format
- Diogo has ICML template from previous work

Format Changes Needed

- Current images not optimized for two-column layout
- Options for image optimization:
    - Condense existing two images into one comprehensive plot
    - Select most impactful plot for main paper, move others to appendix
    - Review additional image from Sharika’s poster for potential inclusion

Content Restructuring

- Current tables and list-style information need reorganization
- Recommendations:
    - Move list-style content to appendix

    - Convert large tables to plots where possible
    - Keep only small, essential tables in main paper
    - Focus on maintaining legibility in two-column format

Workshop Alignment

- Paper already aligns well with new gen workshop focus
- No major justification needed for workshop fit
- Main task is adapting to template requirements
- Project notes contain main points to highlight based on results

Demo Day

- Asked for our paper:
    - Rocio Perales Valdes: http://rvaldes6@gatech.edu/
        - Also asked, “do you have any advice on foundational unlearning papers/resources to start understanding more of the field?”

May 8, 2025

TODOs:

Critical Tasks for Demo Day (May 11th)

~~Technical Preparations~~

1. ~~**Resolve probing accuracy discrepancy**:~~

    - ~~The team reported varying accuracy (40-60% for base model vs target of ~70%)~~
    - ~~Ashwin and Shariqah need to identify what's causing this variation~~
    - ~~Focus only on the highest performing approach (layers -3 and -8 showed 63%)~~

2. ~~**Finalize visualizations**:~~

    - ~~Ensure all figures in the presentation match the latest results~~
    - ~~Verify that the key graphs showing WMDP performance across different unlearning methods properly highlight both:~~
        - ~~The accuracy differences~~
        - ~~The formatting issues explanation~~
    - ~~Add a simple visualization showing how unlearning affects the model's output format~~

~~3. **Prepare demos/examples**:~~

~~- Create 1-2 clear examples showing the same prompt with:~~
~~- Base model response (properly formatted)~~
~~- Unlearned model response (improperly formatted)~~
~~- Rephrased prompt response (properly formatted again)~~
~~- This will visually demonstrate your key finding~~

Presentation Preparation

1. **Assign presentation roles**:

    - Decide who presents (Sharika was mentioned as possible presenter)
    - Ensure they're fully comfortable with all technical details

2. **Streamline narrative**:

    - Restructure slides to emphasize the key discovery about formatting vs. knowledge recovery
    - Ensure the presentation directly addresses how this finding differs from Doshi & Stickland (2024)
    - Practice to fit within 5-minute constraint (approximately 5-7 slides)

Priority Tasks for Workshop Paper (May 19th)

Since the paper is more important and the final report can be adapted from it, these tasks should start immediately after Demo Day:

Technical Analysis

1. **Complete Hindi filler word analysis for ELM**:

    - This appears to be your most novel finding showing recovery of knowledge
    - Run additional tests on ELM models to confirm Hindi filler's effectiveness
    - Analyze why this specific technique works on ELM but not other methods
    - Connect to the meeting notes indicating this was a successful technique

2. **Finalize results on LLM-GAT models**:

    - Complete analysis of all eight unlearning methods
    - Create a comprehensive table showing:
        - WMDP-bio accuracy
        - MMLU/tinyMMLU capability preservation
        - Robustness to different prompting techniques

- Identify clear patterns among methods (the presentation shows three categories)

3. **Probing analysis (if already in progress)**:

    - Prioritize only if Ashwin & Shariqah have made significant progress
    - Focus on showing whether models truly forget knowledge vs. just output formatting
    - Compare with WMDP paper claims about probes not extracting information

Paper Writing

1. **Create detailed paper outline** (4 pages + references):

    - Introduction (~0.5 page): Problem statement, previous work, your contribution
    - Methods (~1 page): Models, datasets, evaluation framework, prompting techniques
    - Results (~1.5 pages): Key findings about formatting issues, method comparisons
    - Discussion (~1 page): Implications, limitations, future work

2. **Draft key sections**:

    - Introduction & Methods: Can be adapted from midterm report
    - Results: Requires updated visualizations and clearer explanation of the formatting discovery
    - Discussion: Needs careful framing of the implications for evaluation methodology

3. **Create high-quality visualizations**:

    - Figure 1: Clear demonstration of formatting issues vs. knowledge recovery
    - Figure 2: Comparison across unlearning methods (highlighting the three categories)
    - Figure 3: Focus on ELM vulnerability to Hindi filler words
    - Table 1: Comprehensive results across methods and models

Additional Tasks (If Time Permits)

1. **Multi-shot analysis**:

    - Run remaining tests on many-shot MMU as mentioned in meeting notes
    - Test bio-retain or bio-forget as few-shot examples

2. **PCA analysis of activations**:

    - Implement using transformer-heads library as mentioned

- Focus on visualizing how knowledge is represented differently after unlearning
- This could provide deeper insight into the formatting issue

3. **Statistical significance testing**:

    - Verify the high scores in the 5% dataset against full dataset
    - Determine confidence intervals for performance differences between methods

Task Distribution (Based on Previous Involvement)

- **Yeonwoo**: Lead on finalizing results analysis, creating visualizations, paper framing
- **Shariqah**: Focus on probing analysis if possible, contribute to paper methods, help with Demo Day
- **Ashwin**: Support probing work, help with technical analysis of results
- **Diogo**: Coordinate overall effort, help with paper writing and review, ensure coherent narrative

Timeline

- **May 8-10**: Focus on Demo Day preparation
- **May 11**: Demo Day
- **May 12-15**: Intensive paper writing and finalization of results
- **May 16**: Paper review and polishing
- **May 17**: Final SPAR report submission (adapted from workshop paper)
- **May 18-19**: Final paper revisions and submission to MUGen

This approach ensures that you prioritize the most important deliverable (the workshop paper) while still meeting the Demo Day requirements, and can easily adapt the paper content for the final SPAR report.

Differences Between Current Draft and Final Workshop Paper

Based on comparing the current draft materials with what's needed for the MUGen workshop submission, here are the key differences:

Structure and Length

The current draft (from your "Eval_for_Unlearning_SPAR_Project.pdf") has the basic structure of a research paper but needs refinement to fit the 4-page ICML workshop format. It appears to be in an earlier draft stage with:

- Adequate abstract and introduction
- Good methodology section but with some redundancy
- Preliminary results that need better organization
- An incomplete discussion section
- Appendix with limited results

Content Gaps

Results Section

1. **Incomplete analysis**: The current draft shows results primarily on RMU with less detail on ELM, TAR, and other methods. The final paper needs a comprehensive comparison across all tested methods.

2. **Missing visualizations**: While there are tables showing RMU results, the paper lacks the key visualizations that appear in your presentation:

    - The 3-category grouping of methods (RMU/PBJ/RR vs. TAR/GradDiff/RepNoise/RMU+LAT vs. ELM)
    - Clear visualization of Hindi filler effectiveness on ELM

3. **Insufficient explanation** of why ELM is vulnerable to certain prompting techniques (especially Hindi filler words) while other methods aren't.

Discussion Section

1. **Underdeveloped implications**: The current discussion doesn't fully explore what your findings mean for the field of unlearning evaluation.

2. **Missing recommendations**: There are no clear guidelines for practitioners on how to better evaluate unlearning methods.

3. **Limited future work**: The paper needs a stronger articulation of what research directions your findings suggest.

Technical Depth

1. **Probing results**: The midterm report and meeting notes mention probing work by Ashwin and Shariqah, but results aren't incorporated into the current draft.

2. **Statistical validation**: The paper doesn't include statistical significance testing of your key findings.

3. **Mechanism exploration**: There's limited discussion of why unlearning affects output formatting rather than actual knowledge.

Format and Polish

1. **ICML formatting**: The current draft doesn't appear to follow the specific ICML template required for the workshop.

2. **Figure quality**: The figures need enhancement for publication quality.

3. **Citation style**: References need to be formatted according to ICML requirements.

4. **Language refinement**: Some sections need tightening for clarity and impact, particularly in highlighting your novel contributions.

Priority Tasks for Completion

To transform the current draft into the final workshop paper:

1. **Restructure results section** to clearly categorize findings across the three groups of unlearning methods

2. **Create publication-quality figures** that effectively demonstrate your key finding about formatting issues vs. knowledge recovery

3. **Develop a stronger discussion** focusing on implications for evaluating unlearning methods

4. **Format according to ICML template** with proper references and citations

5. **Add any new results** from the Hindi filler analysis with ELM and probing work (if completed)

The good news is that you have most of the core content already - the key finding about formatting issues is clearly articulated. The main work involves reorganizing, enhancing visualizations, and strengthening the discussion of implications.

Differences Between Current Draft and Final Workshop Paper

Structure and Length

The current draft has the basic structure of a research paper but needs refinement to fit the 4-page ICML workshop format. It appears to be in an earlier draft stage with:

- Adequate abstract and introduction
- Good methodology section but with some redundancy
- Preliminary results that need better organization
- An incomplete discussion section
- Appendix with limited results

Content Gaps

Results Section

1. **Incomplete analysis**: The current draft shows results primarily on RMU with less detail on ELM, TAR, and other methods. The final paper needs a comprehensive comparison across all tested methods.

2. **Missing visualizations**: While there are tables showing RMU results, the paper lacks the key visualizations that appear in your presentation:

    - The 3-category grouping of methods (RMU/PBJ/RR vs. TAR/GradDiff/RepNoise/RMU+LAT vs. ELM)
    - Clear visualization of Hindi filler effectiveness on ELM

3. **Insufficient explanation** of why ELM is vulnerable to certain prompting techniques (especially Hindi filler words) while other methods aren't.

Discussion Section

1. **Underdeveloped implications**: The current discussion doesn't fully explore what your findings mean for the field of unlearning evaluation.

2. **Missing recommendations**: There are no clear guidelines for practitioners on how to better evaluate unlearning methods.

3. **Limited future work**: The paper needs a stronger articulation of what research directions your findings suggest.

Technical Depth

1. **Probing results**: The midterm report and meeting notes mention probing work by Ashwin and Shariqah, but results aren't incorporated into the current draft.

2. **Statistical validation**: The paper doesn't include statistical significance testing of your key findings.

3. **Mechanism exploration**: There's limited discussion of why unlearning affects output formatting rather than actual knowledge.

Format and Polish

1. **ICML formatting**: The current draft doesn't appear to follow the specific ICML template required for the workshop.

2. **Figure quality**: The figures need enhancement for publication quality.

3. **Citation style**: References need to be formatted according to ICML requirements.

4. **Language refinement**: Some sections need tightening for clarity and impact, particularly in highlighting your novel contributions.

Priority Tasks for Completion

To transform the current draft into the final workshop paper:

1. **Restructure results section** to clearly categorize findings across the three groups of unlearning methods

2. **Create publication-quality figures** that effectively demonstrate your key finding about formatting issues vs. knowledge recovery

3. **Develop a stronger discussion** focusing on implications for evaluating unlearning methods

4. **Format according to ICML template** with proper references and citations

5. **Add any new results** from the Hindi filler analysis with ELM and probing work (if completed)

The good news is that you have most of the core content already - the key finding about formatting issues is clearly articulated. The main work involves reorganizing, enhancing visualizations, and strengthening the discussion of implications.

Apr 30, 2025

Week 12

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	• Played a bit with best-of-N jailbreaking, but doesn't seem amenable to our MCQ setup
Ashwin	☐ Look into probing tutorials ☐ Explore training probes on RMU model. Test probes out on the base models too?	☐	• Worked with Shariqah on training probes for the base zephyr model, but haven't yet matched the probe performance from the wmdp paper (~40% accuracy vs ~70% from the paper)
Shariqah	☐	☐	• Learned about probing • Created a script for probing the baseline zephyr model • Ran evals on addChar for the full wmdp-bio dataset
Yeonwoo	☐	☐	☑ ~~Run the many-shot WMDP-bio formulation, but where the correct answer is passed to the model as being its own previous answer~~ ☐ Check many-shot WMDP-bio, but using bio-retain or

			bio-forget, instead of MCQ

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐ Probing: it would be nice to have an MVP example implemented by Wednesday. Probably the easiest is just a replication of Figure 9 (left) of https://arxiv.org/abs/2403.03218, for RMU and the base model.
        - ☐ Running addChar for the full dataset:
            - ☐ (Nice-to-have) If addChar maintains its performance, check how exactly it is modifying the prompt: is it adding chars at specific places? Specific languages? Etc
        - ☐ Running the LLMU models
        - ☐ (Nice-to-have, but not a must) Run the many-shot MMLU formulation
        - ☐ Run the many-shot WMDP-bio formulation, but where the correct answer is passed to the model as being its own previous answer (instead of coming from the user)
        - ☐ Check many-shot WMDP-bio, but using bio-retain or bio-forget, instead of MCQ (probably worth checking the previous point first)
        - ☐ Not worth pursuing for now:
            - ☐ Testing minor changes in the prompts (i.e. using different models for translation, or rephrasings, etc)

Other (lower priority) Questions:

- Diogo:
    - ☐ We still have a lot of compute left, is there anything that would benefit from more compute?

- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:
        - ☐ There's a **Demo Day on May 11th**, that we are expected to go to (https://spar2025.slack.com/archives/C08CV9DGAKS/p1744319346491939, https://spar2025.slack.com/archives/C08CV9DGAKS/p1745002649360259). We get to present our work publicly there, so we can aim to have the most important aspects completed by then, and then use the last week to tidy everything up.
            - ☐ We need to prepare:

    - A poster
    - A lightning talk
    - (Not required, but we can update the draft as well, since the workshop deadline will be a week afterwards)
- ☐ **SPAR says that Demo Day attendance is required (except for extreme circumstances)**

Pre-meet notes:

- Diogo:

Meeting notes v1

Project Status & Probing Results

- Current probing accuracy results:
    - Base model: 40-60% accuracy (variation between team members)
    - RMU model: ~25% accuracy (random chance)
    - Target from WMD paper: ~70% accuracy for base model
    - Later layers showing better performance (63% for layers -3 and -8)
- Discrepancy in results between team members needs investigation
- Need to verify if seed variations impact results significantly
- Current implementation is working but requires optimization

Demo Day Preparation (Due May 6)

- Deliverables required:
    - Poster
    - 5-minute lightning talk
    - Presentation submission
- Team availability:
    - Sharika: Available Monday-Tuesday only
    - Ashwin: Few hours each day through weekend
    - Yan: Limited availability until Monday-Tuesday
- Assignments:
    - Lightning talk presentation: TBD (Sharika possible presenter)
    - Poster lead: TBD
    - Diogo to potentially lead both if no other volunteers

Technical Investigations & Next Steps

- Many-shot experiments:
    - WMD bio with user-assistant format: similar to previous results
    - Need to run many-shot MMU
    - Testing retain/forget sets as few-shot examples (implementation ongoing)

- Proposed PCA analysis of activations:
    - Could help understand model behavior
    - Potentially show how few-shot examples influence activation patterns
    - Implementation possible using existing transformer-heads library
- Budget status: $50 used out of $500 allocated

Priority Decisions

- Focus for coming week:
    - Primary: Poster and presentation preparation
    - Secondary: Resolve probing accuracy discrepancy
    - If time permits: Run ELM model comparisons
- Timeline considerations:
    - Demo day is on Sunday
    - Submission deadline: May 6
    - Need to allocate significant time for poster/presentation prep
- Team agreed to prioritize demo day deliverables over additional technical investigations

---

Chat with meeting transcript:
https://notes.granola.ai/d/caeaec4f-4add-409a-81e4-0713e64da54b

Meeting notes v2

Project Status Overview

The team discussed their progress on evaluating machine learning model unlearning techniques. The focus has been on testing various methods to retrieve knowledge from unlearned models, with most approaches showing limited success so far. Elm models have shown some promise, while RMU models appear to be more robust against knowledge retrieval attempts.

Upcoming Deadlines

- **Tuesday, May 6th**: Submission deadline for the demo day poster and optional lightning talk
- **Sunday, May 11th**: Demo day where presentations will take place

Technical Updates

Probing Results (Ashwin & Shariqah)

- Working to replicate probing results from the WMDP paper
- Current accuracy for base model probes is around 40-60% (target from paper: ~70%)
- Initial tests on unlearned (RMU) models show random chance accuracy (~25%), which aligns with expectations
- There are discrepancies in accuracy calculations between team members that need resolution
- Probing tests can verify if models retain knowledge independently of formatting issues

Prompting Techniques (Yeonwoo)

- Ran the full dataset using Edar result at Charval
- Results from the 5% subset seem to hold for the full dataset
- Working on many-shot prompting with user-assistant format for WMDP bio
- Exploring the use of retain and forget sets as many-shot examples, but implementation is currently slow

Discussion Points

Project Direction

- Most knowledge retrieval techniques attempted so far are not working well, except potentially with Elm models
- Team considered using PCA analysis on activations to better understand why retrieval techniques fail
- Discussed focusing on framework development as a contribution, even if results are negative

Resource Allocation

- Only $50 spent out of $500 budget so far
- Team has limited time availability before the submission deadline

Decisions

- Primary focus for the next week will be preparing the poster and presentation
- Secondary focus on resolving probing accuracy discrepancies if time permits
- Tertiary focus on running additional tests (many-shot MMLU, activation analysis) if feasible
- Team members will coordinate on Monday/Tuesday to finalize the presentation

---

TODOs for Next Week

Highest Priority (By May 6th)

- [ ] Begin creating the poster for demo day
- [ ] Prepare the lightning talk presentation (5 minutes)
- [ ] Update the project draft with current findings

Medium Priority

- [ ] Resolve probing accuracy discrepancies between implementations (Ashwin & Shariqah)
- [ ] Continue updating the project code and documentation

If Time Permits

- [ ] Run many-shot MMLU tests (Yeonwoo)
- [ ] Test Elm model with normal prompts and successful rephrasing (Diogo suggested)
- [ ] Explore PCA analysis of model activations to visualize unlearning effects
- [ ] Check if the high scores in the 5% dataset are statistically significant

Availability Notes

- Shariqah will be traveling until Monday, May 5th
- Ashwin needs to leave meetings early (at 10:30)
- Most team members have good availability on Monday and Tuesday (May 5-6)

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☑ ~~Identify the discrepancy between the different probe accuracies (on a first glance, I suspect my 40% number is less trustworthy here)~~	☐ Probe results for ELM as well ☐ Probe for the different prompts	

	☐ Poster/presentation?		
Shariqah	☐ poster/presentation	☐ Help Ashwin with probing	
Yeonwoo	☐ Poster/presentation ☐ Run eval w/ LLMU model ☐ Many-shot MMLU ☐ Many-shot bio-retain & bio-forget	☐ addChar analysis 5% vs full set ☐ Probing ☐ Activations + PCA analysis ☐ How sensitive lm-eval results on answer choices (e.g. abcd vs (a)(b)(c)(d), etc.)	

Apr 23, 2025

Week 11

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☐	☐	Unfortunately had to deal with some work issues last week, briefly started looking into probing
Shariqah	☑ ~~Finish implementing https://arxiv.org/pdf/2311.08268~~	☐ Find more simple jailbreak methods and create small datasets	☐ Implemented ReNeLLM mini datasets ☐ Updated the 78 templates datasets so that there was a different dataset per template ☐ Looked for resources/tutorials on linear probing

Yeonwoo	☐ Implement various different jailbreak techniques with increasing complexity	☐ Think about how to evaluate CoT-type answers	☐ Ran evals on n-shot prompts and ReNeLLM ☐ Explored In-Context Attacks as highlighted in https://arxiv.org/pdf/2310.07579

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐ How to deal with many-shot
        - ☐ LLMU: run results

Other (lower priority) Questions:

- Diogo:
- Shariqah:
    - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:
        - ☐

Pre-meet notes:

- Diogo:

Meeting notes v1

Updates from Team Members

Shariqah's Update

- Updated 78 template mini datasets with different complete WMDP samples per template
- Implemented Renee LLM jailbreak rewritings using Gemini
- Investigating linear programming tutorials and resources
- Arena tutorial looks promising but is specific to Ethel GPT
- Checked WMDP repo but could not find their linear probe code
- WMDP uses LM eval for evaluation but doesn't appear to support linear probing

Model Performance Analysis

Zero-Shot vs Few-Shot Results

- Zero-shot performance (~31%) higher than few-shot (~25%) for unlearned models
- Base model maintains capabilities across different shot counts
- Inverted performance curve observed for unlearned models
- Hypothesis: More examples in context may trigger more filtering/degradation
- Need to investigate if model consistently answers specific questions correctly or shows random behavior

Renee LLM Results

- At-char technique showing meaningful bypass of unlearning
- Testing conducted on 5% of total WMDP bio questions
- At-char method: Adding meaningless foreign words without changing semantics
- Performance significantly higher than baseline in LM Gates paper
- Need to verify results across full dataset

Technical Discussion Points

Probing Approach

- Proposal to train probes based on model's activations at each layer
- Would help detect if model internally contains correct information
- Could provide insights into why certain techniques work/fail
- WMDP paper showed probes cannot extract information at any layer
- Could help understand RMU's behavior

Context Learning

- Bio-forget and Bio-retain datasets might work better as context
- Fine-tuning with few examples (10+) shows capability recovery
- In-context learning not showing equivalent results to fine-tuning
- May need to explore MMLU multiple choice questions as alternative context

Next Steps

- Run full dataset evaluation for at-char technique
- Investigate language patterns in successful at-char cases
- Focus on RMU and ELM models for testing
- Implement probing approach (Sherika and Ashwin to coordinate)
- Consider testing MMLU few-shot approach
- Explore alternative translation models for at-char implementation
- Add LLMU testing for all current learning methods

Chat with meeting transcript: https://notes.granola.ai/d/28c60e2f-0adb-4710-947a-2682fc7968f6

Meeting notes v2

Attendees

- Diogo Cruz
- Shariqah Hossain
- Yeonwoo Jang
- Ashwin Sreevatsa

Project Updates

Shariqah's Updates

- Updated the 78 template mini datasets to have different complete WMDP samples per template rather than randomly selecting from all templates
- Implemented a version of the Renalm LLM jailbreak rewritings using Gemini Flashlight
- Explored tutorials for linear programming implementation
- Checked repositories from papers they've been following for linear probe code, but couldn't find WMDP's code for linear probes

Yeonwoo's Updates

- Ran evaluations on prompts including the Renalm techniques that Shariqah implemented
- Found that zero-shot performance was higher than few-shot performance for unlearned models (RMU and ELM), which is the opposite of what was expected
- The "add char" technique from Renalm seems promising for bypassing unlearning, showing meaningful improvement in test results

Key Discussion Points

Unlearning Bypass Techniques

- The "add char" technique (adding meaningless foreign words without changing semantics) showed surprisingly good results compared to other methods
- Results contradict findings from some papers that suggest prompt-based attacks are not effective against unlearning
- Team discussed why this technique might be working and whether it's a statistical fluke or a genuine finding

Model Behavior Analysis

- Discussion about how the unlearned models respond to various prompts and why zero-shot performance is better than few-shot
- Hypothesis from Ashwin: more examples in context might trigger the model's filtering mechanisms, causing it to degrade performance
- Consideration of whether certain models have specific "circuits" that are activated or deactivated during unlearning

Probing as an Evaluation Method

- Discussed using probing to better understand internal model behaviors during unlearning
- Probing could reveal whether models internally know answers but refuse to output them, or truly don't have the knowledge
- Reference to WMDP paper showing that unlearned models genuinely lack the knowledge rather than just refusing to output it

In-Context Learning vs. Fine-Tuning

- Discussion about the difference between in-context learning and fine-tuning for recovering capabilities
- Reference to a paper showing fine-tuning with bio-forget or bio-retain datasets can recover capabilities with just a few examples
- Surprise that in-context learning with examples doesn't yield similar results

TODOs for Next Week

1. **For Shariqah:**
    - Run the "add char" technique on the full dataset to verify if results hold beyond the 5% sample
    - Analyze whether specific foreign languages are more effective in the "add char" approach
    - Continue exploring linear programming for probing implementation
2. **For Yeonwoo & Ashwin:**
    - Coordinate on implementing probing to better understand model internals
    - Test whether different translation models might yield different results for the foreign word insertion techniques
3. **For the Team:**
    - Test the encoding technique where correct answers are formatted as something the model has said in the past

- Consider testing MLU examples as few-shot context instead of WMDP examples
- Run evaluations on LMU model to complete the comparison with other unlearning methods
- Check if providing bio-forget or bio-retain as context (rather than multiple choice questions) helps with recovery

The team is prioritizing understanding why the "add char" technique appears effective and developing more robust evaluation methods through probing.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☐ Look into probing tutorials ☐ Explore training probes on RMU model. Test probes out on the base models too?	☐	
Shariqah	☐	☐	
Yeonwoo	☐	☐	

Apr 16, 2025

Week 10

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	Last week of AI Safety Camp, had to finish that project
Ashwin	☐	☐ Catching up on the last few weeks (midterm	Mostly just taking a pass of the midterm report

		report, meeting notes, etc)	
Shariqah	☑ ~~Look for more easily adoptable jailbreak methods~~ ☐ Create mini datasets for initial testing	☐ Reformat Overleaf paper	☐ Looked at this jailbreak survey: https://arxiv.org/pdf/2407.04295 ☐ Still going through the different papers ☐ implementing rewritten prompts from https://arxiv.org/pdf/2311.08268 ☐ Requested LLMU checkpoints from "Does Unlearning Truly Unlearn" authors
Yeonwoo	☐ Create datasets based on different jailbreak techniques and run evals ☐ Literature review on other prompting techniques ☐ Look into TOFU -> rephrasing as MCQ	☐	☑ ~~Literature review on other prompting techniques~~

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐ Any updates on the jailbreak approach?
        - ☐ LLMU?

Other (lower priority) Questions:

- Diogo:
- Shariqah:
    - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?

    - Diogo:
        - ☐

Pre-meet notes:

- Diogo:

Meeting notes v1

Discussion on Knowledge Retrieval Testing

- Current approach relies on examining logits with LM eval and selecting highest letter
- Limited to first token analysis, which may miss nuanced model knowledge
- Potential weakness: Model doesn't have time to "think through" answers thoroughly

Chain of Thought Implementation Challenges

- Current testing framework not well-suited for chain-of-thought responses
- Issue: First token analysis may not capture full model reasoning
- Model tends to commit to an answer before reasoning through options

Unlearning Process Limitations

- Unlearned models output garbage for certain question types
- Makes it difficult to determine which tokens reflect actual model knowledge
- Challenge in measuring robustness beyond first token responses

Proposed Solutions & Alternatives

- Suggestion to implement prompt engineering with explanation requirements
- Discussed "hacky" approach: Force answer first, then explanation
- Explored scenario nesting as potential workaround
    - May help avoid triggering unlearned prompt responses
    - Could provide alternative logic path for model evaluation

Technical Considerations

- Referenced paper focusing on knowledge retrieval aspects over formatting
- Discussion of multiple choice question analysis methodology
- Need for more robust measuring approach to accommodate various testing techniques

Meeting notes v2

The team discussed progress on their research project focusing on jailbreaking techniques against unlearned language models, specifically testing methods to retrieve knowledge from

models that have undergone unlearning processes. The main discussions centered around potential jailbreaking approaches, evaluation methodologies, and scoping the project for their upcoming workshop paper.

Key Discussion Points

Research Progress and Literature Review

- **Shariqah** shared findings on jailbreak methods from her literature review, identifying two main approaches:
    - **Scenario nesting**: Creating specific contexts for models to respond (e.g., "pretend you're a student in class")
    - **Prompt rewriting**: Reformatting prompts as code, translating portions, etc.
    - She noted implementing a paper with different rewriting techniques and is concerned about applying scenario nesting to WMDP questions which are more knowledge-based than instruction-based.
- **Yeonwoo** identified approximately seven categories of jailbreaking approaches from reviewing 20-30 papers:
    - Template completion, prompt rewriting, scenario nesting
    - Role-playing techniques (e.g., from papers like "Buzz Loomm")
    - In-context attacks with adversarial examples (noting that performance scales with number of examples)
    - Special system tokens injection
    - LLM-based generation of adversarial prompts
    - Relearning through fine-tuning
    - System prompt manipulation

Technical Challenges

- **Diogo** highlighted a key challenge with their evaluation methodology:
    - Current approach looks at the first token's logits to determine which multiple-choice option (A,B,C,D) has the highest probability
    - This approach becomes problematic for jailbreaking techniques that require the model to think through steps before answering
    - The team discussed whether looking at logits throughout a response could be more informative
- **Ashwin** and the team explored possible alternatives:
    - Sampling with higher temperature to detect presence of knowledge
    - Using logits analysis across tokens rather than just the first token

    - Exploring if probes applied to model layers might reveal internal knowledge

Project Scope and Direction

- The team decided to focus on:
    - Prompting-based jailbreaking methods rather than fine-tuning approaches
    - Using logits analysis as their novel contribution
    - Prioritizing techniques that fit within their existing evaluation framework
    - Implementing more straightforward jailbreaks first before exploring complex ones
- **Budget and resources**:
    - Approximately $400 budget remaining
    - 4-5 weeks until project completion
    - Using existing model checkpoints (primarily older models like Zephir) for consistency with literature

Next Week's TODOs

1. **Implement straightforward jailbreak techniques**
    - Focus on prompt rewriting approaches that work with their MCQ evaluation framework
    - Start with simple techniques and progressively explore more complex ones
2. **Prioritize evaluation methodology**
    - Continue using logits analysis as the primary evaluation approach
    - Explore potential ways to generalize the logit reading approach for knowledge retrieval
3. **Literature organization**
    - Consolidate findings from literature reviews
    - Identify which techniques are most promising and most applicable to their WMDP bio task
4. **Define novelty angle**
    - Further develop the logits analysis methodology as the main novel contribution
    - Prepare to frame findings in terms of how they differ from existing literature
5. **Project planning**
    - Update project timeline considering the 4-5 week remaining timeframe
    - Allocate resources efficiently between implementation and paper writing
6. **Documentation**

    - Document results from implemented techniques
    - Begin structuring findings for the workshop paper
    - Fill out next week's goals in the shared document
7. **Communication**

    - Continue discussions in Slack for any questions or clarifications
    - Schedule spontaneous meetings if needed

The team will continue to meet weekly to track progress and adjust priorities as results emerge from their implementations.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☐	☐	
Shariqah	☐ Finish implementing https://arxiv.org/pdf/2311.08268	☐ Find more simple jailbreak methods and create small datasets	
Yeonwoo	☐ Implement various different jailbreak techniques with increasing complexity	☐ Think about how to evaluate CoT-type answers	

Apr 9, 2025

Week 9

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Add Slack updates to meeting notes~~ ☑ ~~Create Overleaf draft~~	☐	

	☑ ~~Do first pass on report~~		
Ashwin	☐	☐	
Shariqah	☑ ~~Follow up on next steps after we have results on other algorithms~~ ☑ ~~Help with midterm report on Sunday, finalize during coworking on Monday~~	☑ ~~Look for non multi-turn jailbreak techniques (LLM-GAT has some techniques)~~ ☐ Look at jailbreaks paper (https://arxiv.org/pdf/2408.15221) for more rephrasing ideas ☑ ~~Create mini datasets (~5% of wmdp) for new rephrasing~~	☐ Submitted midterm report ☐ Created a mini dataset based on https://sites.google.com/view/llm-jailbreak-study/home ☐ Lit review on other template-based jailbreak methods
Yeonwoo	☐ Run more evals on LLM-GAT checkpoints: https://huggingface.co/LLM-GAT ☐ Create matplotlib charts for the midterm report	☐	☑ ~~Run more evals on LLM-GAT checkpoints: https://huggingface.co/LLM-GAT~~ ☑ ~~Create matplotlib charts for the midterm report~~ ☑ ~~Midterm report~~

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
  - Diogo:
    - ☐

Other (lower priority) Questions:

- Diogo:
- Shariqah:
  - [ ] https://arxiv.org/pdf/2310.06987 - can adjust temperature and remove system prompt?

- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
  - Diogo:
    - [ ]

Pre-meet notes:

- Diogo:
  - Jailbreak techniques: https://chatgpt.com/share/67f6a9f7-3a94-8007-aa55-42f25280fca4
    - There are also single-turn ways of mimicking multi-turn jailbreaks, if we'd like to explore that

Meeting notes

Project Status Update

- Team completed and submitted midterm report on Monday
- Ashwin has been absent due to work commitments, expects to return more actively in 1-2 weeks
- Team working on analyzing unlearning methods across multiple models

Technical Results & Analysis

- Completed runs for 8 checkpoints on LLM-GAT
- Current analysis focuses on comparing performance between unlearned vs base models
- Key findings:
  - IndieShiller prompting consistently retrieves more "unlearned" knowledge across all models
  - Most other unlearning methods remain robust
  - Zephyr shows slightly different behavior for Farsi language tests, but team decided this is not significant enough to highlight in report
  - Current evaluation primarily uses WMDP bio and MMLU datasets

Next Steps & Priorities

- Focus on jailbreaking techniques:
  - Will evaluate template-based methods
  - Prioritize simpler, formulaic approaches over complex ones

    - Use 5% WMDP bio dataset for initial testing
    - Consider system prompt modifications
- Potential expansion to TOFU dataset:
    - Has multiple checkpoints for LAMA and FLY
    - Includes built-in evaluation metrics
    - Could be reformatted into multiple choice format

Paper Development Strategy

- Targeting 4-page workshop submission
- Need to:
    - Move content to appendix to meet length requirements
    - Strengthen core message
    - Review recent unlearning papers to ensure novelty
    - Focus on demonstrating robustness evaluation methods
- Team agrees current focus on WMDP/MMLU datasets is sufficient based on previous papers

Task Assignments

- Sharika: Investigate jailbreaking prompts
- Yeonwoo: Explore TOFU dataset implementation
- Ashwin: Review midterm report to catch up
- Team to continue updates via Slack
- Next full team meeting scheduled for following week

Chat with meeting transcript: https://notes.granola.ai/d/cd8ca73f-a224-4a08-b107-0bbeefc1fd9e

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Ashwin	☐	☐ Catching up on the last few weeks (midterm report, meeting notes, etc)	
Shariqah	☐ Look for more easily adoptable jailbreak methods	☐ Reformat Overleaf paper	☐

	☐ Create mini datasets for initial testing		
Yeonwoo	☐ Create datasets based on different jailbreak techniques and run evals ☐ Literature review on other prompting techniques ☐ Look into TOFU -> rephrasing as MCQ	☐	

Apr 7, 2025

Asynchronous

- Diogo: possible workshop:
  - https://mugenworkshop.github.io/ (up to 4 pages) - seems ideal for us, the deadline (May 19th) is at the same line as the end of the program (May 17th)

Apr 2, 2025

Week 8

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐	☐	
Alexander	☐	☐	
Ashwin	☐	☐	
Shariqah	☑ ~~Look at Alex's paper for inspiration on how we can retrieve unlearned info~~ ☑ ~~See if there are improvements we can make to our evaluation~~	☐ Reproduce some results from the paper as needed	- Went through alex's paper and found that the methods either were not focused on knowledge retrieval or were too complex

	~~pipeline design with lm-eval based on how that paper approaches things~~ ☑ ~~Keep team posted on my progress/availability since I am less available first week April~~		to reproduce with our time constraints - Coworking session discussing future rephrasing efforts
Yeonwoo	☐ Reimplement the "truly unlearning" paper with lm-eval and see how the results change ☐ Update the wmdp-bio vs mmlu-overall chart with the lm-eval results	☐	☑ ~~Reimplement the "truly unlearning" paper with lm-eval and see how the results change~~ ☑ ~~Update the wmdp-bio vs mmlu-overall chart with the lm-eval results~~ ☑ ~~Added other unlearned models (Mistral, Llama3) and methods (ELM and TAR)~~

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐

Other (lower priority) Questions:

- Diogo:
    - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:
        - ☐ Write midterm report

Pre-meet notes:

- Diogo:
    - We currently have two directions:
        - Checking the knowledge retrieval results by running more models+methods: Yeonwoo has essentially done that

        - Is there any expectation that we'll find other interesting models publicly available to run?
    - Checking different prompting types: Shariqah is currently doing that - any results so far?

Meeting notes

Key Discussion Points

Team Availability

- **Shariqah**: Limited availability this week due to illness; can work tomorrow, possibly Friday, and Sunday
- **Diogo**: Traveling tomorrow, limited availability; will create Overleaf draft by Friday/Saturday
- **Yeonwoo**: Available to run additional model evaluations; mentioned that they could handle the matplotlib charts

Midterm Report Planning

- **Deadline**: Monday (end of day)
- **Format**: Similar to workshop paper
- **Approach**: Use LLMs for initial draft (Yeonwoo mentioned they could do "95% of the job")
- **Visualization**: Need to include plots showing results; options discussed:
    - Screenshots from Claude's plots (short-term)
    - Creating proper matplotlib visualizations (better approach)
- **Collaboration**: Sunday/Monday collaborations planned for final refinements

Key Research Findings

1. **Formatting Issues**: In reproducing the "Does Learning Trillion" paper, the team discovered that performance drops were mostly due to formatting problems rather than true unlearning
    - Unlearned models often fail to answer in expected ABCD format (~30-40% vs original ~99%)
    - When questions are properly formatted, accuracy remains similar
2. **Method Comparisons**:
    - **RMU**: More robust than ELM
    - **ELM**: Less effective, not as robust as RMU

    - **TAR**: Similar robustness to RMU but shows different capability degradation patterns
3. **New Direction**: Testing additional checkpoints from LLM-GAT, which has eight different unlearning methods

Technical Discussions

- Importance of storing results as CSV files rather than just in README text
- Creating visualizations for comparing performance across methods
- Approaches for running different unlearning method checkpoints

TODOs for Next Week

Diogo

- Add Slack updates to meeting notes
- Create Overleaf draft for the paper
- Do first pass on midterm report

Shariqah

- Follow up on next steps after results on other algorithms
- Help with midterm report on Sunday, finalize during coworking on Monday
- **Stretch goals**:
    - Look for non-multi-turn jailbreak techniques (LLM-GAT has some techniques)
    - Look at jailbreaks paper for more rephrasing ideas
    - Create mini datasets (~5% of WMDP) for new rephrasing

Yeonwoo

- Run more evaluations on LLM-GAT checkpoints: https://huggingface.co/LLM-GAT
- Create matplotlib charts for the midterm report
- Save results in CSV format for easier analysis

Overall Team Priority

The team is focusing on completing the midterm report while continuing to expand testing across different unlearning methods. The report will capture their findings about formatting issues in previous research and the qualitative differences between RMU, ELM, and TAR approaches to unlearning.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐ Add Slack updates to meeting notes ☐ Create Overleaf draft ☐ Do first pass on report	☐	
Alexander	☐	☐	
Ashwin	☐	☐	
Shariqah	☐ Follow up on next steps after we have results on other algorithms ☐ Help with midterm report on Sunday, finalize during coworking on Monday	☐ Look for non multi-turn jailbreak techniques (LLM-GAT has some techniques) ☐ Look at jailbreaks paper (https://arxiv.org/pdf/2408.15221) for more rephrasing ideas ☐ Create mini datasets (~5% of wmdp) for new rephrasing	
Yeonwoo	☐ Run more evals on LLM-GAT checkpoints: https://huggingface.co/LLM-GAT ☐ Create matplotlib charts for the midterm report	☐	

Mar 26, 2025

Week 7

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Check with Alex and Ashwin on availability~~ ☐ Check on charts ☑ ~~Update when2meet~~ ☑ ~~Update Shariqah API setup~~	☐	
Alexander	☐	☐	
Ashwin	☐	☐	
Jan	☐	☐	
Shariqah	☐ Run rephrased prompts on a model other than Zephyr ☐ Try different sizes of Llama (preferably released at the same time so it's controlled) if there are existing unlearned models with that already ☐ Run prompts on other unlearning methods other than RMU	☐ Explore running on WMDP-Cyber	☐ Literature review to find other models and unlearning algorithms that we can run alternate WMDP prompts on ☐ Did not run prompts due to issues with the "Does unlearning truly unlearn" paper ☐ ELM has Zephyr and Llama-3-8B-Instruct on HuggingFace: https://huggingface.co/collections/baulab/elm-6715d68576da0cd1a89c0c04 ☐ RMU on Llama-3-8B-Instruct: https://huggingface.co/collections/baulab/elm-6715d68576da0cd1a89c0c04

Yeonwoo	☐ Add missing rephrasing prompt techniques (i.e. filler) ☐ Chart visualization of the experiment results ☐ Extend evals to cyber tasks ☐ Explore open_unlearning repo and find existing unlearned models to test	☐	☑ ~~Add missing rephrasing prompt techniques (i.e. filler)~~ ☑ ~~5-shots on WMDP~~ ☑ ~~Chart visualization of the experiment results~~ ☑ ~~Found potential flaws in the "truly unlearning" paper~~

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐

Other (lower priority) Questions:

- Diogo:
    - ☐

- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:
        - ☐

Pre-meet notes:

- Diogo:
    -

Meeting notes

In today's meeting, the team discussed important findings regarding the "Does Unlearning Truly Unlearn?" paper and made decisions about next steps for the project. With reduced team capacity (Alex and Ashwin were absent), the focus was on understanding recent discoveries and planning the most efficient path forward.

Key Discussion: Paper Implementation Findings

Diogo and Yeonwoo reported on their analysis of the "Does Unlearning Truly Unlearn?" paper, revealing a significant insight: the paper's conclusions about recovering knowledge from unlearned models may be somewhat misleading.

What they discovered:

- The primary impact of unlearning (particularly RMU) seems to be **formatting issues** rather than true knowledge erasure
- When models are unlearned:
    - They often fail to answer in the expected format (e.g., not using ABCD for multiple choice)
    - Of the questions that are properly formatted (~40% vs original ~99%), the accuracy remains similar to before unlearning
    - The paper's rephrasing techniques (like Hindi translation) primarily help the model return to proper formatting (~90%), not actually recover "forgotten" knowledge
    - The improvement in scores comes from more questions being properly formatted, not from better accuracy on those questions

This finding suggests the paper's conclusion about "recovering knowledge" is potentially misleading - what's being recovered is primarily the ability to properly format responses rather than actual knowledge that was unlearned.

Implementation Plans

The team decided to:

1. Move forward with implementing their approach using lm-eval instead of the paper's setup
    - This will force models to answer regardless of formatting
    - Will provide a better measure of true knowledge retention/unlearning
2. Test Yeonwoo's finding that five-shot prompting doesn't significantly help bypass unlearning
3. Create a 2x2 grid comparing different models and unlearning techniques (as Shariqah suggested)

Project Timeline & Resources

- The team is halfway through the project (Week 7 of ~14)
- A midterm report is due next week
- The team plans to use preliminary plots from Yeonwoo's lm-eval implementation for this report
- Shariqah noted she will have limited availability in the coming week

Decisions on Individual Focus Areas

- **Yeonwoo**: Implement the lm-eval approach to test different prompting techniques properly
- **Shariqah**: Examine Alex's paper for inspiration on knowledge retrieval approaches
- **Diogo**: Coordinate efforts and support implementation

TODOs for Next Week

Yeonwoo:

- Reimplement the "truly unlearning" paper approach with lm-eval
- Update the WMDP-bio vs MMLU-overall chart with the lm-eval results
- Determine if 5-shot prompting genuinely helps bypass unlearning

Shariqah:

- Review Alex's paper for inspiration on knowledge retrieval methods
- Explore potential improvements to the evaluation pipeline based on that paper
- Keep the team updated on availability

Diogo:

- Coordinate between team members
- Prepare for midterm report
- Review any preliminary results
- Support implementation efforts as needed

The team's primary goal is to complete the LM-eval implementation and run initial experiments to have results for the midterm report.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑	☐	
Alexander	☐	☐	

Ashwin	☐	☐	
Jan	☐	☐	
Shariqah	☑ ~~Look at Alex's paper for inspiration on how we can retrieve unlearned info~~ ☑ ~~See if there are improvements we can make to our evaluation pipeline design with lm-eval based on how that paper approaches things~~ ☑ ~~Keep team posted on my progress/availability since I am less available first week April~~	☐ Reproduce some results from the paper as needed	
Yeonwoo	☐ Reimplement the "truly unlearning" paper with lm-eval and see how the results change ☐ Update the wmdp-bio vs mmlu-overall chart with the lm-eval results	☐	

Mar 19, 2025

Week 6

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Set up Runpod and a Claude API key~~ ☐ Look into getting Meeting notes during the meeting ☑ ~~Get Alex updates~~ ☑ ~~Check on updates at the end of the week~~	☐ Draw flowchart with future steps for the project	

Alexander	☐	☐	
Ashwin	☐ Run MMLU, WMDP using the techniques in this paper on Zephyr-7B for base, unlearned models to get datapoints for chart	☐ Clean up implementation	
Jan	☐	☐	
Shariqah	☑ ~~Sync with Yeonwoo on rephrasing prompts on Friday~~ ☑ ~~Follow up with authors to see if they have scripts/WMDP data~~ ☐ Implement rephrasing for WMDP	☐ Get familiar with lm-eval	☐ Wrote initial scripts for rephrasing with some minimal testing, but seems like I may have overlapped with Yeonwoo's work? ☐ No response from authors about WMDP rephrase datasets
Yeonwoo	☑ ~~Get to Goal 1 (run full evals on MMLU and WMDP for both base and unlearned models including prompt rephrasing)~~	☐ Rewrite the code using lm-eval	☑ ~~Get to Goal 1 (run full evals on MMLU and WMDP for both base and unlearned models including prompt rephrasing)~~

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
    - Diogo:
        - ☐ How's the implementation of WMDP+RMU with different prompts (i.e. Goal 1)?

Other (lower priority) Questions:

- Diogo:
    - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
    - Diogo:
        - ☐

Pre-meet notes:

- Diogo:
    - The team currently has reduced capacity:
        - Jan unfortunately has other commitments, and will take a more passive role on the team
        - Alex might have a bit of time before the ICML review period, but he might be too busy for a while. As a result, it's best to consider his tasks nice-to-haves
        - Ashwin expects to be quite busy until end of March, at least
    - As a result, and since we are currently in Week 6, we might have to focus on the core tasks, and potentially downsize the scope of the project (until we have a better clue of how far we can get)
    - A possible MVP would be doing a sort of extension of the "Does Unlearning truly unlearn?" paper, with the associated lm-eval implementation. For that, we would need to get the Goal 1 results, for various prompts, and see what obvious easy extensions we can make
        - We should probably avoid extensions that require finetuning new models (since we might not have the manpower for it)
        - We might focus on low hanging fruit that the original paper missed
            - Especially approaches that are fast to implement

Meeting notes

Attendance and Updates

The meeting had limited attendance, with only Diogo Cruz, Yeonwoo Jang, and Shariqah Hossain present. Diogo explained that several team members have reduced availability:

- Jan has too many commitments and will take a more passive role
- Alexander might have limited time before ICML review period
- Ashwin is expected to be busy until at least the end of March

Project Resources Discussion

- Yeonwoo mentioned running low on RunPod credits (only $3 left)
- Diogo offered to top up the credit and set up separate API keys to track individual spending
- Both RunPod and Vast.ai resources are available for team use

Technical Progress Updates

Yeonwoo reported on implementing various prompting techniques for evaluating unlearning:

- Successfully ran experiments including prompt rephrasing tests
- Results saved in the "erasure" branch with documentation in the README
- Used Claude Haiku instead of Opus to reduce costs

- Found that the current RMU model setup has issues with formatting answers correctly
- Approximately 30-40% of questions weren't answered in the expected ABCD format

Evaluation Discussion

- The team discussed the difference between their current evaluation approach and LM-eval
- LM-eval looks at logits for ABCD answers while their current approach checks if the model outputs a valid answer format
- There might be significant differences in results between these methods
- Yeonwoo plans to create visualization charts of the experiment results

Project Scope Reconsideration

Given the reduced team capacity, Diogo suggested possibly downsizing the project scope:

- Focus on core tasks and create a "Does Unlearning Truly Unlearn?" paper extension
- Prioritize techniques that are fast to implement and don't require fine-tuning new models
- Concentrate on low-hanging fruit the original paper might have missed

Next Steps Discussion

The team discussed several directions:

1. Create visualization charts of current results
2. Add missing prompt techniques (like "filler" words)
3. Potentially extend evaluation to cybersecurity tasks
4. Look for other unlearning methods beyond RMU to compare against
5. Consider testing with different model sizes to see if unlearning robustness depends on scale
6. Explore the idea of a 2D grid comparing different unlearning techniques and models

Team Dynamics and Scheduling

- Discussion about co-working sessions and their utility given the team's changed availability
- Agreement to keep co-working sessions but potentially reschedule them for better attendance
- Concerns raised about whether absent team members can effectively catch up later
- Diogo will clarify expected availability with Alex and Ashwin

Conclusion

The meeting concluded with an agreement to focus on completing the current evaluation framework, visualizing results, and potentially expanding to a 2x2 grid of models and unlearning techniques if feasible.

TODOs for Next Week

Diogo

- Check with Alex and Ashwin regarding their availability
- Look at existing charts and help create visualizations
- Update the when2meet link for potentially rescheduling co-working sessions
- Top up RunPod credits for the team
- Set up separate API keys for tracking individual usage

Shariqah

- Run rephrased prompts on models other than Zephyr
- Try different sizes of Llama models (preferably released at the same time for controlled comparison)
- Run prompts on unlearning methods other than RMU
- Explore running on WMDP-Cyber (stretch goal)

Yeonwoo

- Add missing rephrasing prompt techniques (specifically the "filler" method)
- Create chart visualizations of the experiment results
- Extend evaluations to cybersecurity tasks if possible
- Explore the open_unlearning repository to find existing unlearned models to test

Alexander & Ashwin

- (Pending clarification from Diogo on their availability)

Next Week Goals

	Goal	Stretch goal	What actually happened?

Diogo	☑ ~~Check with Alex and Ashwin on availability~~ ☐ Check on charts ☑ ~~Update when2meet~~ ☑ ~~Update Shariqah API setup~~	☐	
Alexander	☐	☐	
Ashwin	☐	☐	
Jan	☐	☐	
Shariqah	☐ Run rephrased prompts on a model other than Zephyr ☐ Try different sizes of Llama (preferably released at the same time so it's controlled) if there are existing unlearned models with that already ☐ Run prompts on other unlearning methods other than RMU	☐ Explore running on WMDP-Cyber	
Yeonwoo	☐ Add missing rephrasing prompt techniques (i.e. filler) ☐ Chart visualization of the experiment results ㅜㅅㄷrac ☐ Extend evals to cyber tasks ☐ Explore open_unlearning repo and find existing unlearned models to test	☐	

Mar 12, 2025

Week 5

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Update Ashwin, Jan~~ ☑ ~~(I'll complete the rest tomorrow)~~	☐	
Alexander	☐ Check the code for prefix optimization / system prompt / … from Florian paper(https://arxiv.org/abs/2409.18025) ☐ Implement prefix optimization in embedding space instead of GCG	☐	
Ashwin	☐	☐	Implemented a basic version of the "modifying multiple-choice responses" prompts from "LLM Unlearning Benchmarks are weak measures of progress" paper, though the paper itself did not seem to have a github so haven't confirmed if this is exactly the same implementation
Jan	☑ ~~Methods Overview on 'Robustness Measures'~~ ☐ Adversarial prompts ☐ quantization attacks	☐ Sync with Shariqah, on WMDP prompts	Busy with conference hosting and paper submission No sync, overview not fully completed - want to tie more to project direction https://github.com/diogo-cruz/eval_for_unlearning/tree/erasure
Shariqah	☐ Get rephrased versions of WMDP prompts based on [2411.12103] Does Unlearning Truly Unlearn? A Black Box Evaluation of LLM Unlearning Methods	☐ Get familiar with lm-eval	Not as much progress as I hoped because wasn't feeling well. Synced a bit with Yeonwoo. Reached out to the authors to see if they can share WMDP rephrasings

	☑ ~~Reach out to authors or created prompts from scratch as needed~~ ☐ Run MMLU rephrased prompts as a starting point		
Yeonwoo	☐ ~~Build on lm-eval and make the pipeline more efficient / robust (e.g. incorporating different prompting techniques)~~	☐	Ran WMDP eval on the Zephyr-RMU model using the "Does unlearning truly unlearn" paper code – 0-shot result matches with the paper but there's some issues with 5-shot implementation; plan to work on rephrasing as the next step

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
  - Diogo:
    - ☐ How's the implementation of WMDP+RMU with different prompts?
    - ☐ How much does the OpenUnlearning repo change our approach?

    Alexander:
    - ☐ Should I help with implementic class wrapper for different few-shot examples?

Other (lower priority) Questions:

- Diogo:
  - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
  - Diogo:
    - ☐

Pre-meet notes:

- Diogo:
  -

Meeting notes

The meeting began with Diogo Cruz confirming that Alexander would not be attending due to a seminar commitment. After team members completed updates in the shared document, they proceeded with a round of updates from each person.

Key Updates

Jan Batzner

Jan discussed his literature review work, focusing on a Nature Machine Intelligence paper that provides a good overview of the field. He's been analyzing what additional value can be added to the project, particularly regarding robustness evaluations. He mentioned his interest in hearing about the current project direction to help determine the most relevant criteria to focus on.

Project Direction Discussion

Diogo explained that the literature review and coding aspects of the project have become intertwined, with team members implementing papers they've been reviewing. He highlighted that the project currently has three main paper implementations:

1. The paper Alexander is following
2. The paper Ashwin is working on
3. The "Does Unlearning Truly Unlearn?" paper that Shariqah and Yeonwoo are implementing

Diogo suggested that the team now has sufficient starting points and should focus more on the coding implementation, which he identified as the current bottleneck. He mentioned that the literature review can be revisited later when preparing the related work section for their paper.

Yeonwoo Jang

Yeonwoo provided detailed updates on implementing the "Learning Truly Unlearn" paper:

- Successfully replicated the paper's zero-shot results on the WMDP dataset
- Encountered issues with answer formats, noting that about 30-40% of questions went unanswered because the model didn't produce responses in the expected ABCD format
- Found discrepancies in the five-shot implementation, where the authors use MMLU examples for prompting WMDP questions, which seemed counterintuitive
- Observed that five-shot prompting increased the percentage of questions answered but slightly degraded performance

Yeonwoo also shared a relevant GitHub repository, OpenUnlearning, which provides a framework for different unlearning techniques and includes pre-trained models that could be useful for the project.

Shariqah Hossain

Shariqah reported limited progress due to illness but mentioned:

- Reached out to paper authors requesting their rephrasing scripts
- Planning to work on implementing rephrasing techniques using OpenAI's API if necessary
- Working with MMLU datasets provided in CSV format

Ashwin Sreevatsa

Ashwin discussed his work implementing techniques from the "LLM Unlearning Benchmarks are Weak Measures of Progress" paper:

- Implemented an approach that replaces one incorrect multiple-choice answer with a term related to data that should be unlearned (like "SARS COVID-19")
- Discussed the possibility of integrating this into LM-eval for testing
- Working on a basic implementation that could be connected to evaluation frameworks

Technical Discussions

Several technical discussions occurred throughout the meeting:

1. **Evaluation Methods**: The team discussed differences between LM-eval and the current implementation, particularly how they handle multiple-choice questions
2. **Data Formats**: Discussions about how different papers handle answer formats and question ordering
3. **Implementation Strategy**: Debate about whether to prioritize getting results quickly versus standardizing implementation approaches
4. **OpenUnlearning Repository**: Discussion about how this repository might influence their approach, with Diogo suggesting it could be a useful resource but might shift focus toward robustness testing

Computing Resources

Yeonwoo suggested moving from Fast AI to RunPod for more stable computing resources and requested setting up a Claude API key for rephrasing tasks. Diogo explained the reimbursement process for compute costs and agreed to set up the requested resources.

Next Steps and Timeline

Diogo addressed project timeline, noting they had initially planned for a presentation this week but were behind schedule with no tangible results yet. He mentioned upcoming deadlines:

- A midterm report for SPAR in about 2-3 weeks
- An initial draft deadline in week 10
- Team members estimated they could have initial results by next week's meeting

Closing

The meeting concluded with Diogo offering to join a co-working session after the meeting for those interested and mentioning he would update the meeting notes later when the transcript became available.

TODOs for Next Week

Diogo

- Set up RunPod and a Claude API key for the team
- Look into getting meeting notes during the meeting (rather than relying on delayed transcripts)
- Check with Alexander for updates
- Check on progress updates at the end of the week
- Draw flowchart with future steps for the project (stretch goal)

Ashwin

- Run MMLU and WMDP using techniques from the "LLM Unlearning Benchmarks" paper on Zephyr-7B for both base and unlearned models to get datapoints for a comparison chart
- Clean up implementation (stretch goal)

Shariqah

- Sync with Yeonwoo on rephrasing prompts on Friday
- Follow up with authors to see if they have scripts/WMDP data
- Implement rephrasing for WMDP
- Get familiar with lm-eval (stretch goal)

Yeonwoo

- Get to Goal 1: run full evaluations on MMLU and WMDP for both base and unlearned models including prompt rephrasing
- Rewrite the code using lm-eval (stretch goal)

Alexander

- (To be confirmed upon return)

Jan

- (To be confirmed based on upcoming literature review direction)

The team's primary focus for next week is to generate tangible results showing how different prompting approaches affect both base and unlearned models' performance on MMLU and WMDP benchmarks.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Set up Runpod and a Claude API key~~ ☐ Look into getting Meeting notes during the meeting ☐ Get Alex updates ☐ Check on updates at the end of the week	☐ Draw flowchart with future steps for the project	
Alexander	☐	☐	
Ashwin	☐ Run MMLU, WMDP using the techniques in this paper on Zephyr-7B for base, unlearned models to get datapoints for chart	☐ Clean up implementation	
Jan	☐	☐	
Shariqah	☐ Sync with Yeonwoo on rephrasing prompts on Friday ☐ Follow up with authors to see if they have scripts/WMDP data ☐ Implement rephrasing for WMDP	☐ Get familiar with lm-eval	
Yeonwoo	☐ Get to Goal 1 (run full evals on MMLU and WMDP for both	☐ Rewrite the code using lm-eval	

	base and unlearned models including prompt rephrasing)		

Mar 5, 2025

Week 4

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Coordinate with Shariqah~~ ☑ ~~Converge on implementation on Friday with Yeonwoo~~ ☑ ~~Continue lit review~~	☐	
Alexander	☑ ~~Try Zephyr_RMU & Zephyr-7b-alpha on MMLU & wmdp splits (**before Friday** https://huggingface.co/cais/Zephyr_RMU & also mixtral)~~ ☑ ~~Check how one can pre-fill with different prompts~~	☐ Check Florians paper for metrics (https://arxiv.org/abs/2409.18025)	
Ashwin	☐ Work with yeonwoojangus@gmail.com as needed for testing eval for new prompts	☐ (Look through papers in Meta-o1-pro subtab)	
Jan	☑ ~~Review Meta-o1-pro subtab (Section 5.2; Shariqah TOFU) diogo.abc.cruz@gmail.com and shariqah97@gmail.com s work, expand on it~~ ☑ ~~Expand on Min 2024 table~~	☑ ~~Prioritize papers in Meta-o1-pro subtab for review, make time plan~~	**Next week** -Methods Overview on 'Robustness Measures' - Adversarial prompts - quantization attacks

Shariqah	☑ ~~Coordinate with Diogo and Yeonwoo on code~~ ☑ ~~Start looking at alternate prompting~~	☐	- Started running code from [2411.12103] Does Unlearning Truly Unlearn? A Black Box Evaluation of LLM Unlearning Methods - Doesn't provide WMDP prompts in repo
Yeonwoo	☐ Build on Alexander's eval implementation to test different prompts (e.g. WMDP)	☑ ~~Work on another implementation using Inspect in parallel~~	- Implemented an end-to-end eval pipeline using Inspect and discovered that it's not ideal for MCQ

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
  - Diogo:
    - ☐ How's the implementation of WMDP+RMU with different prompts?

    Alexander:
    - ☐ Should I help with implementic class wrapper for different few-shot examples?

Other (lower priority) Questions:

- Diogo:
  - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
  - Diogo:
    - ☐

Pre-meet notes:

- Diogo:
  -

Meeting notes

The meeting began with team members Alexander Panfilov and Shariqah Hossain providing their updates, while awaiting other team members who were delayed. Yeonwoo Jang joined later, with Ashwin and Jan unable to attend due to scheduling conflicts.

Key Updates

Alexander Panfilov

- Successfully reproduced the ZRMU model results from the MDP paper in his code
- Explored how to implement a wrapper for models to test different prompting techniques
- Investigated using a prompt from the "Does Learning Include Unlearn" paper to elicit unlearning behavior
- Also looked into Florian's paper regarding prefix optimization techniques

Shariqah Hossain

- Investigated the "Does Unlearning Truly Unlearn" paper and got the code running
- Faced challenges with the WMDP prompts, as they weren't provided in the repository due to claimed access requirements
- Considering creating her own rephrased prompts or reaching out to the paper authors
- Noted that the paper doesn't clearly show how they generated their altered prompts

Yeonwoo Jang

- Compared evaluation approaches between LM-eval and Inspect
- Found that LM-eval works better for multiple-choice questions as it examines log probabilities for answer choices
- In contrast, Inspect requires strict adherence to answer format templates
- Recommended using LM-eval for MCQ-type benchmarks given its more flexible evaluation approach

Key Discussion Points

1. **Prompt-Based Evaluation**: The team discussed how unlearned models behave with normal prompts, noting that unlearning often causes models to output "garbage" when faced with prompts they shouldn't respond to, rather than responding with refusals or "I don't know."

2. **Evaluation Framework Decision**: The team agreed to use LM-eval for now as it's better suited for multiple-choice questions, which are easier to evaluate and will allow for faster iteration.

3. **Future Evaluation Extensions**: While starting with MCQ evaluation, the team acknowledged the need to eventually include more realistic setups with open-ended responses and tools like ROUGE scoring.

4. **Task Distribution**:

    - Alexander will focus on prefix optimization from Florian's paper

- Shariqah will work on getting rephrased WMDP prompts
- Yeonwoo will improve the LM-eval pipeline for testing different prompting techniques

Next Steps

The team assigned specific goals for the coming week and agreed to push changes to the main branch to prevent the development branches from diverging too much.

TODOs for Next Week (By Team Member)

Alexander

- Check the code for prefix optimization/system prompt from Florian's paper (https://arxiv.org/abs/2409.18025)
- Implement prefix optimization in embedding space instead of using GCG (Gradient-based Controlled Generation)

Ashwin

- Tasks to be determined upon return (not present at meeting)

Diogo

- Update Ashwin and Jan (who were absent) on meeting outcomes
- Complete remaining TODOs (to be specified tomorrow)

Jan

- Create a methods overview on 'Robustness Measures'
- Research adversarial prompts and quantization attacks

Shariqah

- Obtain rephrased versions of WMDP prompts based on the "Does Unlearning Truly Unlearn?" paper
- Either reach out to paper authors or create prompts from scratch if necessary
- Run MMLU rephrased prompts as a starting point
- Get familiar with lm-eval as a stretch goal

Yeonwoo

- Build on lm-eval to make the pipeline more efficient and robust
- Incorporate support for different prompting techniques in the evaluation framework

The team expects to have a better understanding of how different prompting techniques affect unlearned models by the next meeting, with progress on both evaluation framework improvements and specific unlearning tests.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☐ Update Ashwin, Jan ☐ (I'll complete the rest tomorrow)	☐	
Alexander	☐ Check the code for prefix optimization / system prompt / … from Florian paper(https://arxiv.org/abs/2409.18025) ☐ Implement prefix optimization in embedding space instead of GCG	☐	
Ashwin	☐	☐	
Jan	☐ Methods Overview on 'Robustness Measures' ☐ Adversarial prompts ☐ quantization attacks	☐ Sync with Shariqah, on WMDP prompts	
Shariqah	☐ Get rephrased versions of WMDP prompts based on [2411.12103] Does Unlearning Truly Unlearn? A Black Box Evaluation of LLM Unlearning Methods ☐ Reach out to authors or created prompts from scratch as needed	☐ Get familiar with lm-eval	

	☐ Run MMLU rephrased prompts as a starting point		
Yeonwoo	☐ Build on lm-eval and make the pipeline more efficient / robust (e.g. incorporating different prompting techniques)	☐	

Mar 3, 2025

Coworking

- Diogo:
  - ☐ Quick discussion with Yeonwoo
    - Apparently, the base and unlearned models don't always respect the template answer provided in the prompt when outputting their multiple choice answer
    - In particular, for the default Inspect multiple_choice implementation, WMDP gets around 0.3-0.4/0.13 for the base/unlearned models, instead of 0.6/0.3.
    - Lm_eval has a much more permissible approach to parsing the model's answer, so it always (?) assigns a letter to the answer, even if the model outputs garbage.
  - ☐ Quick discussion with Shariqah
    - General updates
    - Focusing on understanding the "different prompts" repo

Feb 27, 2025

Coworking

- Diogo:
  - ☐ Quick discussion with Shariqah

Feb 26, 2025

Week 3

Last Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Continue lit review~~ ☑ ~~Clarify its direction with Jan and Shariqah~~ ☑ ~~Clear up time availability with Alex~~ ☑ ~~Help set up cloud instance and pipeline~~	☐ Finish lit review	
Alexander	☑ ~~Implement small eval replication (e.g. WMDP or TOFU) using model checkpoints~~	☐	Drafted eval on lm-eval-harness tasks (evaled wmdp on TAR-model & llama2_7b)
Ashwin	☑ ~~Looking into cloud instance setup for evaluations~~ ☑ ~~Get a basic demo of evaluation script on existing unlearned checkpoints~~	☐	Tested out WMDP eval on GPT2, Zephyr 7B
Jan	☐ Identify impactful papers in the community (nodes) ☑ ~~Summarize limitations and opportunities for future research~~	☐	Want to expand on the limitations, e.g. make overview table like in Min 2024 p. 4
Shariqah	☐ Touch base with Jan and Diogo for lit review ☐ Lit review with focus on benchmarks	☐	
Yeonwoo	☑ ~~Cloud instance setup for local runs~~	☐	Wrote a script for running eval end-to-end using the Inspect library that can fetch

	☑ ~~Implement small eval replication (e.g. WMDP or TOFU) using model checkpoints~~		dataset and base model from huggingface and run a simple eval (tested on WMDP & llama-3.2-1B-Instruct)

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
  - Diogo:
- Other (lower priority) Questions:
  - Diogo:
    - ☐
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
  - Diogo:
    - ☐

Pre-meet notes:

- Diogo:
  - We should discuss what is in the notes from the last coworking session

Meeting notes

The team met to discuss progress on their machine unlearning evaluation project, sharing updates and planning next steps. Here's a summary of what was covered:

Updates from Team Members

- **Alexander**: Implemented evaluation using lm-eval-harness for WMDP on TAR-model and llama2_7b
- **Ashwin**: Tested WMDP evaluation on GPT2 and Zephyr 7B models
- **Jan**: Working on literature review, identifying impactful papers and summarizing limitations of current approaches
- **Yeonwoo**: Set up cloud instance, wrote script for running end-to-end evaluation using the Inspect library on WMDP and llama-3.2-1B-Instruct

Key Discussion Points

1. **Project Direction**: The team agreed to focus on the WMDP benchmark as a starting point, particularly looking at biosecurity/cybersecurity unlearning tasks using RMU (Reinforcement-based Machine Unlearning).

2. **Immediate Goal**: Establish a two-part plan:

    - **Goal 0**: Replicate the WMDP benchmark paper to ensure the evaluation pipeline works
    - **Goal 1**: Test how different prompts affect the ability to extract supposedly "unlearned" information

3. **Literature Review Insight**: Diogo discussed a paper called "Does Unlearning Truly Unlearn?" that demonstrates how unlearned models can still recall information when prompted differently - exactly what the team plans to investigate.

4. **Compute Resources**: The project has a specific credit card for compute costs, and team members should use this for any work requiring payment.

5. **Task Distribution**: The team identified the need to test using existing unlearned model checkpoints rather than training models from scratch, with Alexander volunteering to work on a simple replication before Friday.

Plan for Next Week

The team assigned the following goals:

- **Diogo**: Coordinate with Shariqah, converge on implementation with Yeonwoo, continue literature review
- **Alexander**: Test Zephyr_RMU & Zephyr-7b-alpha on MMLU & WMDP splits before Friday
- **Ashwin**: Work with Yeonwoo for testing evaluation with new prompts
- **Jan**: Review literature in Meta-o1-pro subtab (particularly Section 5.2) and expand on Min 2024 table
- **Yeonwoo**: Build on Alexander's evaluation implementation to test different prompts

TODOs for Next Week

1. **Alexander**: Create a basic implementation of evaluation using lm-eval-harness on unlearned models (deadline: before Friday)

    - Test Zephyr_RMU and Zephyr-7b-alpha on MMLU & WMDP tasks
    - Check how to implement different prompts in the evaluation

2. **Yeonwoo & Ashwin**: Build on Alexander's implementation to test different prompts

    - Focus on the WMDP dataset
    - Explore techniques from adversarial papers to elicit supposedly unlearned information

3. **Jan**: Continue literature review

- Review papers in Meta-01-pro subtab, particularly Section 5.2
- Expand on Min 2024 table for limitations of current approaches

4. **Diogo**: Coordinate overall efforts

    - Work with Shariqah on literature review direction
    - Help converge implementation approaches by Friday

5. **All team members**: Document progress in the shared repository

    - Push code changes to the GitHub repo for sharing
    - Update meeting notes with any significant findings

The team is working toward creating a visualization similar to the one shown in the meeting notes, but with different points representing different prompts rather than different hyperparameters, to demonstrate how robust (or not) various unlearning approaches actually are.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Coordinate with Shariqah~~ ☐ Converge on implementation on Friday with Yeonwoo ☐ Continue lit review	☐	
Alexander	☑ ~~Try Zephyr_RMU & Zephyr-7b-alpha on MMLU & wmdp splits (**before Friday** https://huggingface.co/cais/Zephyr_RMU & also mixtral)~~ ☑ ~~Check how one can pre-fill with different prompts~~	☐ Check Florians paper for metrics (https://arxiv.org/abs/2409.18025)	
Ashwin	☐ Work with yeonwoojangus@gmail.com as needed for testing eval for new prompts	☐ (Look through papers in Meta-o1-pro subtab)	
Jan	☑ ~~Review Meta-o1-pro subtab (Section 5.2; Shariqah TOFU)~~	☑ ~~Prioritize papers in Meta-o1-pro subtab for~~	**Next week** -Methods Overview

	~~diogo.abc.cruz@gmail.com and shariqah97@gmail.com s work, expand on it~~ ☑ ~~Expand on Min 2024 table~~	~~review, make time plan~~	on 'Robustness Measures' - Adversarial prompts - quantization attacks
Shariqah	☐	☐	
Yeonwoo	☐ Build on Alexander's eval implementation to test different prompts (e.g. WMDP)	☐ Work on another implementation using Inspect in parallel	

Feb 24, 2025

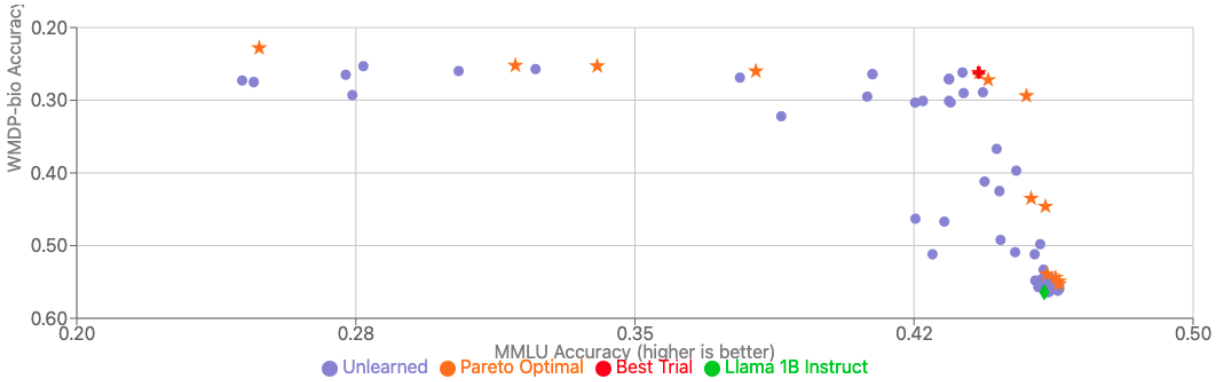
Coworking

- Diogo:
  - ☐ Quick discussion with Ashwin about replication setup
    - GPT-2 is not good enough, since it gets poor MMLU scores even in the base setting
    - Llama 1B instruct is probably a better alternative (though Zephyr-7B is probably good enough to start with, and more robust)
  - ☐ Discussion with Shariqah about Lit review
    - Writing notes in main **Literature review** tab
  - ☐ Checking existing unlearned models
    - https://huggingface.co/locuslab
    - https://huggingface.co/OPTML-Group
    - Unfortunately these are not available through Huggingface's Inference API, so we need to run them locally
  - ☐ **Potentially good starting point. Focus on:**
    - Unlearning task: Biosecurity/Cybersecurity -> proxy: WMDP-Bio/Cyber
    - Method: RMU (starting point), later other techniques
    - Model: either Zephyr-7B, or a smaller model (like Llama 1B)
    - Test different prompts like in https://arxiv.org/abs/2411.12103
    - Output: a plot of MMLU vs WMDP-Bio/Cyber results, for different prompt approaches. Something like this (but for different prompt approaches):

MMLU vs WMDP-bio Accuracy with Pareto Frontier

Higher MMLU and Lower WMDP-bio are better

Unlearned versions of Llama 1B model, compared to random chance (0.25) and Llama 1B instruct



Analysis Summary

The scatter plot shows the relationship between MMLU accuracy (higher is better) and WMDP-bio accuracy (lower is better). The Pareto frontier (orange line) connects points that represent optimal trade-offs between these two metrics.

The best trial according to the optimization objective (highlighted in red) has:

- MMLU accuracy: 0.4424 (higher is better)
- WMDP-bio accuracy: 0.2620 (lower is better)
- Objective value: 0.2772

Key reference points:

- Random-chance baseline is 0.25 for both metrics (dashed gray lines)
- Llama 1B instruct model (green diamond): MMLU = 0.460, WMDP-bio = 0.564

This plot shows unlearned versions of the Llama 1B model. The best trial demonstrates a good trade-off by achieving substantially better-than-random MMLU performance (0.442 vs 0.25 random) while keeping WMDP-bio accuracy much lower than the fully learned Llama 1B instruct model (0.262 vs 0.564).

This suggests the optimization was successful in finding model configurations that maintain reasonable knowledge abilities while reducing unwanted benchmark performance.

Feb 20, 2025

Coworking

- Diogo:

Feb 19, 2025

Coworking

- Diogo:
  - [x] ~~Writing pre-meeting notes~~
  - [ ] Helped set up vast.ai instance
    - Yeonwoo and Ashwin have used it before
    - If it turns out to not be a good option, we can just move to another solution
  - [ ] Added instructions about replication to Resources tab
  - [ ] Updated meeting notes (4o -> o1 pro)

Week 2

Pre-meeting questions

Write things here throughout the week or while we wait for everyone to arrive at the meeting.

- What are the ~3 **most important questions** to answer in this meeting?
  - Diogo:
    - [ ] What are your major uncertainties right now?
    - [ ] Benchmark and metrics:
      - [ ] How much can we reuse from https://github.com/EleutherAI/lm-evaluation-harness or https://github.com/UKGovernmentBEIS/inspect_ai, versus doing it from scratch?
    - [ ] How accessible are unlearned models? Can you use existing models through an API or do we need to train them from scratch?
- Other (lower priority) Questions:
  - Diogo:
    - [ ] Any difficulties you keep running into?
- Right now, what do we think are the **~3 most important things (MITs)** to achieve over the next week?
  - Diogo:
    - [ ] Get an MVP implementation of a task (e.g. WMDP + RMU unlearned model)
    - [ ] What is the optimal setup to start with:
      - [ ] Run locally with API model access
      - [ ] Run a cloud instance

Pre-meet notes:

- Diogo:
    - I haven't had a meeting with Jan this week, so the lit review is currently WIP
        - Are any of the other groups bottlenecked by info they need from the lit review?
- I was thinking that there are two immediate directions:
    - Taking an immediate unlearning task (like cybersecurity) and making it more robust. For this case, we could just start with unlearning cybersecurity, and:
        - Implement the pipeline to evaluate it (if it doesn't exist out-of-the-box)
        - Check existing unlearning metrics for unlearning cybersecurity specifically
        - Develop more robust metrics, building on the simple evals we started with
    - Building a framework that can take a general task, and evaluate it.
        - It would require a great deal of automation:
            - Automate generating the evals (though maybe they could be provided as an optional argument)
        - It is probably too demanding to start with. In practice, most interest seems to be on a few unlearning tasks (like cybersecurity).

Meeting notes

- **Weekly Meeting Format**: Diogo reiterated the structure:
    - *Pre-meeting questions* (capture topics in the shared doc)
    - *Round of updates* from each person
    - *Discussion of current pain points and next steps*
    - *Action items (TODOs) for next week*
- **Literature Review**:
    - Jan could not attend; thus, progress on the literature review (especially around the most common benchmarks/tasks used in unlearning) is still pending.
    - Some of the implementation details (e.g. which tasks/data sets to focus on first) may depend on forthcoming clarifications from the lit review.
- **Benchmarking & Data Sets**
    - *Unlearned models via API?*
        1. Yeonwoo found no straightforward public API for pre-trained *unlearned* models. Likely we must **run or fine-tune them locally** (on a cloud instance with GPU if needed).
    - *Data sets and code bases examined so far*:
        1. **WMDP (Withdrawal of Memorized Data in Pretrained Models)**
            - The QA portion is easy to access from Hugging Face.
            - The corpora for biology or other subfields sometimes require special approval (or local loading to avoid Google Drive blocks).
            - The standard metrics are mostly accuracy-based, but the group wants to check how robust that is for unlearning evaluations.
        2. **Tofu**
            - Tofu provides unlearned checkpoints for multiple baselines (on Llama, Pythia, etc.).

        - Sharika and Ashwin have scripts in a new `benchmark/` branch to load data and compare Tofu's QA style to WMDP's multiple-choice tasks.
    - *Metrics/Eval Tools*
        1. **lm-evaluation-harness** (from EleutherAI) vs. **Inspect** (from UK AIS)
            - *LMEvalHarness* has more off-the-shelf tasks (including some WMDP tasks), is easier to set up.
            - *Inspect* is more flexible with logs/visualization but can require more customization.
- **Immediate Direction**
    - The group converged on doing an **MVP "replication style" pipeline** first:
        1. Take an existing unlearned checkpoint (e.g. from WMDP, Tofu, or another open-source source).
        2. Evaluate it on the relevant QA or multiple-choice tasks.
        3. Verify that we can replicate the reported results (e.g. that the unlearned model truly "forgets" targeted knowledge).
    - Longer-term, they can expand to more robust or custom metrics, but starting with an established benchmark helps avoid "building from scratch" and ensures having a working end-to-end pipeline.

Action Items (TODOs) for Next Week

1. **Set Up a Computing Environment**
    - Since there are no ready-made APIs, the team must run unlearned models locally or in the cloud. Diogo will help set up a GPU-powered cloud instance and show how to run these large models (e.g. Tofu's unlearned checkpoints).
2. **Implement a Simple MVP Evaluation**
    - **Pick one unlearned checkpoint** (e.g. from Tofu or WMDP) and one of the corresponding tasks (cybersecurity or biology QA, etc.).
    - **Load the model + data** in either `lm-evaluation-harness` or a simple custom script.
    - **Check performance** against baseline metrics (e.g. accuracy).
    - Document the steps so the team can iterate or extend to more tasks.
3. **Coordinate with Literature Review**
    - Once Jan's lit review clarifies which benchmarks are used most consistently in unlearning research, the team can finalize which tasks/data sets to support first.
    - The group will revisit whether a "starter subset" of data or the full WMDP/Tofu sets are ideal for next expansions.
4. **Continue Exploratory Scripting**
    - Sharika & Ashwin will keep refining the "benchmark/" branch scripts for WMDP and Tofu data.

    - Yeonwoo will compare the configurability of `lm-evaluation-harness` vs. `Inspect`, especially for unlearning metrics.
5. **Maintain Ongoing Questions in the Shared Doc**
    - Use the “Pre-meeting questions” section during the week to note discussion topics, uncertainties, or roadblocks that arise before the next meeting.

Key Goal for Next Meeting:
Have at least one small end-to-end run (model + QA data + baseline metric) completed in a cloud environment, so the team can confirm the pipeline works and begin iterating on more robust metrics or data sets.

Next Week Goals

	Goal	Stretch goal	What actually happened?
Diogo	☑ ~~Continue lit review~~ ☑ ~~Clarify its direction with Jan and Shariqah~~ ☑ ~~Clear up time availability with Alex~~ ☑ ~~Help set up cloud instance and pipeline~~	☐ Finish lit review	
Alexander	☐	☐	
Ashwin	☑ ~~Looking into cloud instance setup for evaluations~~ ☐ Get a basic demo of evaluation script on existing unlearned checkpoints	☐	
Jan	☐	☐	
Shariqah	☐ Touch base with Jan and Diogo for lit review ☐ Lit review with focus on benchmarks	☐	
Yeonwoo	☐ Cloud instance setup for local runs ☐ Implement small eval replication (e.g. WMDP or	☐	

	TOFU) using model checkpoints		

Feb 17, 2025

Coworking

- Diogo:
  - ☐ Clarifying pipeline with Yeonwoo and Alex

Feb 14, 2025

Coworking

- Diogo:
  - ☐ Going through the lit review TODOs from yesterday
  - ☐ Helping out with any blockers

Feb 13, 2025

Asynchronous

- Diogo (list of things that came to mind):
  - Lit review TODOs (some of these overlap with other groups):
    - Gather list of unlearning techniques (and associated refs and repos, if available)
      - We may focus on SOTA techniques, that the field currently cares about
    - Gather list of testing techniques (benchmarks, probing, fine-tuning, etc), and associated repos
    - Add all these to **Resources** tab
  - Group M TODOs:
    - List of possible ways to evaluate unlearned models. E.g.:
      - Finetune the base model ourselves on a cloud instance and evaluate it there
      - Find known unlearned models accessible through API and evaluate those locally (through API)

    - If feasible, this would be preferable as a starting point, since it makes this simpler and faster to iterate

Feb 12, 2025

Kick-off meeting

Meeting Plan (3 hours total)

Hour 1: Team Introduction and Ice-Breaking (60 mins)

- First 15 mins: Quick round of personal introductions
- 45 mins: Zoom session for 1-on-1s
    - Each team member should meet others individually
    - With 6 team members total, everyone should get to meet 5 others
    - ~8-9 minutes per pair
    - Round 1: ⏳ 8:00
    - Round 2: ⌛ 0:00
    - Round 3: ⌛ 0:00
    - Round 4: ⌛ 0:00
    - Round 5: ⌛ 0:00

Break (5 mins)

Hour 2: Project Overview and Discussion (50 mins)

- 10 mins: Project overview by Diogo
    - Review key goals from proposal
    - Explain the three initial groups (3 groups)
- 20 mins: Open discussion about the project
    - Team members can share their perspectives and ideas
    - Discuss initial group preferences (already collected in the doc)
- 20 mins: Discussion of potential failure modes and mitigation strategies

Break (5 mins)

Hour 3: Project Practicalities (60 mins)

- 15 mins: Weekly meeting scheduling
    - Review everyone's time zones (UTC-8 to UTC+1)
    - Use the when2meet already created to find optimal recurring meeting time
- 20 mins: Project setup and tools
    - Set up team Slack channel

        - Review Google Drive structure
        - Discuss compute resources (vast.ai plan)
        - Github repository setup
- 15 mins: Next steps and immediate tasks
    - Confirm initial group assignments
    - Set concrete goals for each group's first week
    - Ensure everyone has clear next steps

Meeting notes

-

Goals

- For code-specific TODOs, I recommend you use Github issues.

🛠️ Resources

Include any relevant, recurring useful links or resources here.

Time availability:

- **Preferred:** https://www.when2meet.com/?28913334-Vb2ec
- **Extended:** https://www.when2meet.com/?28913349-4qkE6

Paper with nice figures: https://arxiv.org/pdf/2311.04046

Midterm report: SPAR Midterm Report

Template: [Make a Copy] SPAR Midterm Report Template

Overleaf: https://www.overleaf.com/2114818437rvjwxfhzqbrq#713c14

WMDP replication links:

- Github: https://github.com/jeanne-s/wmdp_rec
- Doc: Research Engineers Guide
- Instructions: Setting up instance

Vast.ai instance instructions:

- What you do in the instance would either be easily downloadable (i.e. models) or should be committed to the Github repo
- You should have a setup script to easily setup the environment in a new instance
- If you're working on the instance outside of coworking hours, send a message in Slack, otherwise Diogo might assume that you forgot to turn off the instance and turn it off himself

Unlearning Taxonomy: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10880482

Budget notes: Accessing SPAR compute

Demo Day:

- **Presentation:** SPAR Demo Day presentation
- **Poster:** Demo Day poster

⏪ Literature review

Notes

- Diogo:
    - Currently going through:
        - [2410.16454] Catastrophic Failure of LLM Unlearning via Quantization
        - [2411.12103] Does Unlearning Truly Unlearn? A Black Box Evaluation of LLM Unlearning Methods
            - Seems like a good starting point for our project, they covered WMDP-bio with RMU and LLMU
            - They have a repo: https://github.com/JaiDoshi/Knowledge-Erasure
- Shariqah
    - https://arxiv.org/pdf/2502.05209
        - an interesting expansion on using model tampering rather than just prompting for evaluating unlearning
    - https://arxiv.org/abs/2402.16835 Eight Methods to Evaluate Robust Unlearning in LLMs
        - No code
        - Focuses on *"Who's Harry Potter? Approximate Unlearning in LLMs."*
            - "based on fine-tuning using text that has
            - been modified to replace domain-specific content with generic content"
            - '"Familiarity" metric, which is designed to measure the model's ability to complete Harry Potter content as determined by an automated GPT-4 evaluation'
        - "we find i) higher-than-baseline amounts of knowledge can reliably be extracted, ii) WHP performs on par with the original model on Harry Potter Q&A tasks, iii) it represents latent knowledge comparably to the original model, and iv) there is collateral unlearning in related domains. Overall, our results highlight the importance of comprehensive unlearning evaluation that avoids ad-hoc metrics"
        - Evaluation
            - Tested unlearned prompts on other languages - showed unlearning generalized to other languages
            - Jailbreaking the unlearned info
                - leads to modest increases in the WHP model's Familiarity
                - Based on work in this repo: https://github.com/verazuo/jailbreak_llms
            - ICL w unlearned Harry Potter context - increased familiarity
            - Few-shot fine tuning to try to resurface unlearned info
                - Significantly increased familiarity
            - Evaluate on trivia questions about Harry Potter

        - ”tasks require question-answering behavior as opposed to the… text generation that was directly unlearned by the WHP method”
        - unlearned model performed well, indicating poor unlearning
    - Check for unlearned info remaining in latent knowledge - supervised linear and unsupervised contrastive probes
        - Able to retrieve answer for easier prompts
        - Here is code from “Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness?“ which was referenced for linear probes: github.com/lingo-mit/lm-truthfulness
    - Attempted to unlearn via prompting - unsuccessful
    - “find that WHP loses significant Familiarity in related domains, including English Mythology and Harry Potter film production”
- “Unveiling Entity-Level Unlearning for Large Language Models: A Comprehensive Analysis” https://aclanthology.org/2025.coling-main.358.pdf
    - No code
    - “erase entity-related knowledge from the target model” as opposed to “instance-level unlearning, specifically targeting the removal of predefined instances containing sensitive content”
    - “current methods struggle to achieve effec- tive entity-level unlearning… the knowledge coverage of the forget set and its size play pivotal roles…entities introduced through fine-tuning are more vulnerable than pre-trained entities during unlearning”
    - Forget set construction: “prompt the target models to self-generate entity-related questions according to their internal knowledge”
    - Baseline unlearning algorithms
        - Gradient Ascent
        - Gradient Difference
        - KL Minimization
        - Preference Optimization
        - Negative Preference Optimization
    - Evaluation metrics - uses TOFU benchmark
        - ROUGE
        - “Probability computes the conditional probabil ity of QA pairs in the evaluation set”
        - Accuracy “calculates the proportion of a para- phrased answer that the unlearned model can select from perturbed answers of the question”
        - Forget Quality and model utility as defined in TOFU
        - “Knowledge Cov- erage, designed to assess the knowledge overlap between the constructed forget sets and the target set. This metric is computed using the BERTScore (Zhang et al., 2019)

of the closest QA pair match between the constructed forget set and the target set. “
        - Also “we construct forget sets with varying degrees of coverage by systematically re- placing different ratios of QA pairs within the con- structed forget set with those from the target set”
    - performance of the algorithms on constructed forget sets improves with size
- ”knowledge introduced during fine-tuning is more vulnerable to unlearn ing interventions” - more damage to out of scope knowledge

- Jan: Adding onto this
    - Robustness Measures besides Lynch 2024
        - **MUSE (Shi et al., 2024)**: Six-way evaluation measuring verbatim and paraphrased memorization, privacy leakage, general capability, scalability, and multi-step unlearning https://arxiv.org/html/2407.06460v1
        - **TOFU's Metric Suite (Maini et al., 2024)**: Uses Truth Ratio and "p-value for forget quality" comparing to a model that never saw the data https://arxiv.org/abs/2401.06121
    - “Quantization Attacks” (Zhang et al. 2024) https://arxiv.org/abs/2411.12103
        - Quantization (reducing model precision) can sometimes restore supposedly unlearned knowledge
        - This suggests that unlearning may only suppress information at the full-precision level
        - When the model is compressed or quantized, the suppression mechanism might break down
        - This represents a serious vulnerability for deployment scenarios where models are often quantized for efficiency

- Jan:

Related work	Unlearning targets/tasks	Influence erasure methods	**Effectiveness**: **(I)** In-scope evaluation for unlearning efficacy **(O)** Out-of-scope evaluation for model utility	Efficiency
(Lu et al., 2022)	Reducing toxic content, avoiding undesirable sentiments, and preventing repeated text generation	Reward-reinforced model fine-tuning	**(I)** Toxic prompts, specific sentiments, & repetitive sentences **(O)** Unlearning target-irrelevant prompts	N/A
(Jang et al., 2022)	Degenerating private information, w/ unlearning response irrelevant to this info	Gradient ascent-based fine-tuning	**(I)** Prompts from training data extraction **(O)** Natural language understanding tasks	Runtime cost
(Kumar et al., 2022)	Text de-classification, w/ unlearning response close to that of retraining*	Sharded, isolated, sliced, and aggregated (SISA) training via adapter	**(I)** No evaluation for unlearning efficacy **(O)** Test set	Runtime cost Memory cost
(Ilharco et al., 2022) (Zhang et al., 2023c)	Degenerating toxic content	Task vector-based parameter-efficient fine-tuning via LoRA	**(I)** Prompts leading to toxic generation **(O)** Perplexity on other datasets	N/A
(Wang et al., 2023)	Text de-classification/de-generation, unlearning specific words in translation, w/ response close to that of retraining*	KL-divergence-based fine-tuning	**(I)** Training subset **(O)** Test set	Runtime cost
(Yu et al., 2023)	Unlearning gender and profession bias, with de-biased unlearning response	Weight importance-informed & relabeling-based fine-tuning	**(I)** Biased prompts **(O)** No evaluation for model utility	N/A
(Pawelczyk et al., 2023)	Text de-classification, w/ unlearning response close to that of retraining*	In-context learning	**(I)** Training subset **(O)** Retain & test sets	Black-box access
(Eldan & Russinovich, 2023)	Degenerating Harry Potter-related book content, w/ unlearning response irrelevant to Harry Potter	Relabeling-based fine-tuning	**(I)** Questions and their rephrased/hard versions about Harry Potter **(O)** NLU tasks	N/A
(Ishibashi & Shimodaira, 2023)	Unlearning knowledge from QA dataset, with refusal response (*e.g.*, 'I don't know')	Relabeling-based fine-tuning	**(I)** Adversarial and original questions about forgotten knowledge **(O)** Other QA prompts	N/A
(Chen & Yang, 2023)	Text de-classification and de-generation, with response close to that of retraining*	KL divergence-based parameter- efficient fine-tuning via adapter	**(I)** Training subset **(O)** Retain & test sets	Runtime cost
(Wu et al., 2023b)	Degenerating private information, w/ unlearning response irrelevant to this info	Importance-based neuron editing	**(I)** Memorized private data points **(O)** Test set	Runtime cost
(Yao et al., 2023)	Degenerating harmful prompts, degenerating Harry Potter-related book content, and reducing hallucination	Integration of gradient ascent, random labeling, & KL divergence-based fine-tuning	**(I)** Prompts related to unlearning targets **(O)** NLU tasks	Runtime cost
(Maini et al., 2024)	TOFU: Unlearning biographical knowledge about fictitious authors	Fine-tuning with various objectives	**(I)** Q&A about the unlearning authors **(O)** Q&A about other authors and general facts	Runtime cost
(Patil et al., 2024)	Degenerating sensitive information using factual information as a testbed	Model editing techniques and constrained finetuning	**(I)** Prompts for unlearned factual knowledge **(O)** Prompts for unrelated factual knowledge	White-box v. black-box access
(Thaker et al., 2024)	Harry Potter questions and author biography in TOFU (Maini et al., 2024)	Guardrailing with a separate LLM	**(I)** Q&A about Harry Potter and unlearning authors **(O)** Standard NLP benchmarks	N/A
(Zhang et al., 2024b)	Fictitious unlearning using TOFU (Maini et al., 2024)	Negative preference optimization	Same as TOFU (Maini et al., 2024)	N/A
(Li et al., 2024a)	Hazardous knowledge in the domain of biology, cybersecurity, and chemistry	Optimization towards random representations for unlearning concept	**(I)** Zero-shot Q&A about hazardous knowledge **(O)** Zero-shot Q&A about other general knowledge, and fluency of models	N/A
(Barbulescu & Triantafillou, 2024)	Specific text sequences memorized by LLM	Memorization-aware gradient ascent	**(I)** Memorization scores of the forget samples **(O)** Commonsense and scientific reasoning tasks	N/A
(Wang et al., 2024c)	Private, toxic, and copyrighted knowledge	Factual relation removal in MLP layers	**(I)** Accuracy of generating ground-truth knowledge **(O)** Evaluation on reasoning abilities	N/A
(Wang et al., 2024a)	Fictitious unlearning using TOFU (Maini et al., 2024)	Reverse KL divergence based knowledge distillation	**(I)** Q&A about the unlearning authors **(O)** Commonsense and scientific reasoning tasks	N/A
(Liu et al., 2024)	Fictitious unlearning using TOFU (Maini et al., 2024), hazardous knowledge using WMDP (Li et al., 2024a), copyrighted content in news articles and book	Detecting the forget prompts and corrupting their embedding space	**(I)** Q&A or completion of the unlearned knowledge **(O)** Eleven common LLM benchmarks	Runtime cost

(Liu et al., 2024, CoRR arXiv version, p. 4)

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C.Y., Xu, X., Li, H. and Varshney, K.R., 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp.1-14. (linked in proposal)

1. Concept: seperate unlearning into four components: unlearning targets, influence erasure, effectiveness, and efficiency.

2. Method: They categorize approaches into model-based (parameter modification) and input-based (instruction design) methods, noting that most research prioritizes model-based approaches like gradient ascent variants and localization-informed techniques.

3. Critical Gaps identified:

3.1 Data-model interactions that require joint examination

3.2 Connections with model editing technique

3.3 Potential for adversarial training to enhance robustness

3.4 Applications of reinforcement learning principles to unlearning

3.5 Challenges of continual unlearning during model lifecycles

4. **Assessment Framework**

4.1 Effectiveness: comparing with retraining, testing "hard" in-scope examples, and robustness against attacks

4.2 Utility preservation: maintaining capabilities outside unlearning scope

4.3 Efficiency: computational cost, memory requirements, and scalability

5. Future Research: (most interesting for us)

5.1 Examine sensitivity to hyperparameters, comp. resources, data-model select.

5.2 Prioritize robustness against jailbreaking & relearning; certified unlearning
5.3 Develop standardized benchmarks and optimization techniques balancing effectiveness with utility
5.4 Investigate interpretability methods (saliency maps, example-based explanations, loss landscapes)
5.5 Establish governance frameworks for unlearning practices, addressing privacy, security, and fairness
5.6 Create "LLM Unlearning Algorithm Cards" documenting parties, data, evaluations, and implementation details

Adding from Liu et al. 2025 (Nature M.Int.):

- Methodology Overview for Robustness: Lynch, A., Guo, P., Ewart, A., Casper, S. & Hadfield-Menell, D. (2024) - "Eight Methods to Evaluate Robust Unlearning in LLMs"
- Quantization for Unlearning: Zhang, Z. et al. (2024) - "Does Your LLM Truly Unlearn? An Embarrassingly Simple Approach to Recover Unlearned Knowledge"
- Adversarial: Łucki, J. et al. (2024) - "An Adversarial Perspective on Machine Unlearning for AI Safety"
- Fan et al. (2024) - "Challenging Forgets: Unveiling the Worst-Case Forget Sets" https://arxiv.org/html/2403.07362v1

Next:

- Yuan, X., Pang, T., Du, C., Chen, K., Zhang, W. and Lin, M., 2024. A Closer Look at Machine Unlearning for Large Language Models. *arXiv preprint arXiv:2410.08109*. (20 graph connection)
- Gao, H., Pang, T., Du, C., Hu, T., Deng, Z. and Lin, M., 2024. Meta-Unlearning on Diffusion Models: Preventing Relearning Unlearned Concepts. *arXiv preprint arXiv:2410.12777*. (15 graph connections)
- Hu, S., Fu, Y., Wu, Z.S. and Smith, V., 2024. Jogging the Memory of Unlearned LLMs Through Targeted Relearning Attacks. *arXiv preprint arXiv:2406.13356*. (15 graph connections)

Meta-o1-pro

Below is a **comprehensive, single-document literature review** that collates the main points from the multiple smaller reviews above. This integrates references, explains key findings, and contextualizes them for building a **robust evaluation framework** for LLM unlearning.

Literature Review: Machine Unlearning in Large Language Models (2024–2025)

Machine *unlearning* in Large Language Models (LLMs) seeks to remove specific knowledge or behaviors that a model has acquired—such as proprietary text, privacy-violating data, harmful content, or factual errors—*without* retraining entirely from scratch. This capability is increasingly important for meeting regulatory "right to be forgotten" requests, mitigating harmful or copyrighted content, and ensuring user privacy. Yet **robustly** verifying that an LLM has truly forgotten targeted data or behavior is difficult. Below, we survey the last ~\sim12–18 months of work on unlearning methods, benchmarks, and especially **evaluation** techniques, highlighting strengths and persistent gaps.

---

1. Unlearning Methods in LLMs

1.1 Fine-Tuning with "Negative" Objectives

A straightforward approach is to *fine-tune the model* on the data to be forgotten but *reverse* or *penalize* correct predictions on that data (a "negative training" procedure).

- **LLMU** (Yao et al., 2023) and **Gradient Ascent** (Jang et al., 2023)
    - Apply *gradient ascent* on the forget set: maximizing loss so that the model *performs poorly* on those inputs.
    - They often combine this with a "retain set" or a KL regularization step to preserve other knowledge.
    - Empirically effective at raising perplexity on the forget data, but can produce side-effects if tuned poorly.
- **RMU** (Li et al., 2024), **LLMU** vs **RMU** comparison (Doshi & Stickland, 2024)
    - **RMU** updates only certain layers or submodules while "negating" the memory for the target set.

    - Tends to preserve general performance better than a naive full gradient-ascent approach but can still be circumvented by adversarial prompts or “prompt engineering.”

1.2 Parameter-Efficient Approaches

Given that full fine-tuning is expensive (especially for multi-billion-parameter models), several works propose:

- **LoRA-Based Unlearning** (Chen et al., 2024; Gundavarapu et al., 2024)
    - Insert small adapter/LoRA modules trained to *counteract* the forbidden data.
    - Achieves partial unlearning while updating only a small fraction of parameters, typically leading to fewer side effects.
- **Adapter “Unlearning Layers”** (Chen & Yang, 2023, “EUL”)
    - Introduce dedicated “unlearning layers” that are finetuned for forgetting.
    - Allows multiple unlearning requests by merging or fusing separate adapters, each forgetting different data.

1.3 Model Editing and Knowledge Negation

- **Mechanistic Unlearning** or “Weight Surgery” approaches locate and modify the parameters that encode specific facts or behaviors (Guo et al., 2024; Kim et al., 2024).
    - Potentially faster (since changes are localized) but can be fragile if knowledge is widely distributed.
- **NegMerge**, “Weight Negation” (Kim et al., 2024)
    - Directly modifies key parameter subsets to “cancel out” knowledge.
    - Shows strong forgetting but can degrade other capabilities and is vulnerable to adversarial reactivation.

1.4 Distillation and Data Deletion

- **Distillation-Based** unlearning can remove malicious “backdoors” or narrow capabilities by forcing a student model to *ignore* the forbidden content from a teacher model (2410.14425).
- **Data Deletion** techniques (Zhang et al., 2024, “Negative Preference Optimization”) actively push the model *away* from the memory of targeted training data.

1.5 Multi-Objective and Continual Unlearning

- **MoLLM** (Pan et al., 2024)
    - Frames unlearning as a multi-objective optimization: forgetting the target while retaining other knowledge.
    - Shows improved stability compared to naive gradient ascent alone.
- **Sequential or Continual Unlearning**
    - Real deployments might require repeated removal of new data or knowledge.
    - Early works handle this by combining small adapter-based unlearning steps over time, but robust solutions remain open research.

Key Takeaway:
Most current methods revolve around *fine-tuning with reversed objectives*, *parameter-efficient patches*, or *localized weight editing*, often plus a "retain set" to reduce collateral damage. While these approaches are promising, **none** offers complete forgetting without side effects, and they remain vulnerable to *adversarial revival* of knowledge.

---

2. Evaluation Challenges and Failure Modes

Despite quick progress, verifying unlearning has proven difficult. Recent studies highlight **common failure modes**:

1. **Residual Memory / Recoverable Knowledge**
    - Methods sometimes only *suppress* knowledge. Prompting in a different style, using multi-step queries, or fine-tuning again can re-activate forgotten info (Doshi & Stickland, 2024; Zhang et al., 2024 on quantization).
2. **Over-Forgetting / Collateral Damage**
    - Tuning for strong forgetting often destroys unrelated knowledge or degrades general performance. This trade-off is central to every approach (Yao et al., 2024).
3. **Refusal vs. True Unlearning**
    - Models might output blanket refusals or disclaimers to queries about the forgotten data—rather than *genuinely lacking* the knowledge (Ma et al., 2025; Yang et al., 2024). Distinguishing that from real forgetting is tricky.
4. **Adversarial Attacks**
    - Clever prompts ("jailbreaks"), partial re-training, or model manipulations (e.g., quantization) can *undo* or reveal the unlearned info (Lynch et al., 2024).

5. **Dependency & Entanglement**

    - Knowledge is often entangled in multiple tokens or contexts. Removing one piece can lead to partial or incomplete forgetting or inadvertently removing related knowledge (Wu et al., 2024 on “deep unlearning”).

Hence, robust **evaluation** must test multiple angles: direct queries, paraphrases, adversarial prompting, partial retrieval, “retain set” performance, etc.

---

3. Benchmarks for Unlearning

In the last year, **public benchmark datasets** emerged to standardize testing:

1. **WMDP** (Li et al., 2024)

    - *Weapons of Mass Destruction Proxy* with 3,668 harmful knowledge QA pairs.
    - Evaluates if a model can forget malicious instructions while preserving other knowledge.
2. **TOFU** (Maini et al., 2024)

    - *Task of Fictitious Unlearning*, with 200 synthetic author profiles.
    - The model must forget a subset of these synthetic individuals’ details, testing how well it “behaves as if it never saw them.”
    - Their “Truth Ratio” metric compares the unlearned model to a *clean* model trained without those profiles, exposing partial forgetting.
3. **EDU-RELAT** or **Synthetic Fact Datasets** (Wu et al., 2024)

    - Synthetic knowledge bases track logical dependencies among facts.
    - Tests “deep unlearning” by seeing if the model not only forgets a fact but also the *inferred* facts.
4. **Harmful/Backdoor Sets**

    - Datasets for removing malicious triggers or harmful QAs (Gundavarapu et al., 2024; Kim et al., 2024).
    - Evaluate how effectively the model ceases producing harmful completions.
5. **Entity-Level Data** (Ma et al., 2025; Yang et al., 2024)

    - Test removing all knowledge about a single person or entity.
    - Evaluate how that affects queries about similar or related entities.

Key patterns:

- Benchmarks typically provide a *"forget set"* plus a *"retain set"*, checking the model's drop on the forget set vs. retained performance.
- Some incorporate adversarial prompts or paraphrases.
- Studies like **TOFU** demonstrate that *no existing approach* fully mimics a truly "clean" model.

---

4. Evaluation Metrics and Frameworks

Evaluating unlearning means measuring:

1. **Forgetting Completeness**
    - *Accuracy or perplexity on the "forget set"* should drop drastically.
    - *Exposure or membership inference tests* (Carlini et al.) measure if the model still leaks memorized strings.
    - *Truth Ratio* (TOFU) compares unlearned model outputs to a *clean model* (which never saw the data).
2. **Utility / Capability Preservation**
    - *Accuracy on a retain set*, or perplexity on general tasks.
    - Harmonic means or multi-objective scores that penalize big drops in general performance.
    - Some works use *factual correctness checks* on general QA tasks.
3. **Robustness / Adversarial**
    - Testing re-phrasings, multi-step prompts, or *jailbreak attacks* to see if the data can be resurrected (Lynch et al., 2024).
    - Checking if a small fine-tune can re-learn the forgotten info suspiciously quickly.
4. **Collateral Damage**
    - Metrics for how the model's performance changes on data *similar* to the forget set, or on other entities from the same fine-tuning batch (Ma et al., 2025).
5. **Model Output Quality**
    - *Token Diversity, semantic coherence, factual correctness* in unaffected domains (Yuan et al., 2025).
    - A model that "forgets" by refusing or giving nonsense is not a success; the evaluation must detect that.

Notable Proposed Frameworks

- **MUSE** (Shi et al., 2024)
    - Outlines a “six-way evaluation”: measuring verbatim and paraphrased memorization, privacy leakage, general capability, scalability, and multi-step unlearning.
- **Eight-Method Evaluation** (Lynch et al., 2024)
    - Emphasizes adversarial prompt tests, cross-lingual attempts, re-learning speed, hidden-state analysis, etc.
    - Shows how a seemingly thorough unlearning (e.g., “Who’s Harry Potter?” approach) fails under more robust testing.
- **TOFU’s Metric Suite** (Maini et al., 2024)
    - Uses *Truth Ratio* and a “p-value for forget quality” by comparing to a model that never saw the data.
    - Also aggregates multiple utility metrics into a single “Model Utility” score.
- **Xu et al. (2024) Survey**
    - Categorizes evaluation methods: *time-based*, *accuracy-based*, *similarity to a retrained model*, and *attack-based* (membership inference, model inversion).
    - Concludes that no single metric suffices.

Common Themes:
(a) Compare the unlearned model to a *clean “retrained from scratch”* model (the “gold standard”), or at least approximate one.
(b) Use multiple metrics capturing forgetting efficacy, utility retention, and adversarial robustness.
(c) Evaluate not only numeric accuracy but also the model’s *behavior and output style* to detect refusal vs. real forgetting.

---

5. Comparative Insights from Recent Studies

1. **TOFU Baseline Studies (Maini et al., 2024)**
    - Showed that all tested methods still left traces of the forget data (the unlearned model is distinguishable from a truly clean model).
    - Some methods forget *too aggressively*, hurting general performance.
2. **Does Unlearning Truly Unlearn? (Doshi & Stickland, 2024; Zhang et al., 2024)**
    - Highlighted easy “prompt-based reactivation” and “quantization attacks” that restore knowledge.
    - Emphasize the need for adversarial checks.

3. **SKU for Harmful Knowledge (Liu et al., 2024)**
    - Achieves lower *harmful response rate* while preserving normal perplexity.
    - Demonstrates unlearning can be improved if metrics evaluate safe vs. normal prompts distinctly.
4. **Harry Potter (Eldan & Russinovich, 2023) vs. Lynch et al. (2024)**
    - “Who’s Harry Potter?” method seemed to remove a large domain.
    - But deeper testing found partial leftover knowledge. Showcases that a single test (e.g. perplexity) might be insufficient.
5. **Entity-Level Unlearning (Ma et al., 2025; Yang et al., 2024)**
    - Fine-tuning-introduced knowledge is more fragile and can be over-deleted or partially linger.
    - Methods like preference optimization sometimes yield universal refusals (not real forgetting).

Conclusions:

- Current methods do not achieve bulletproof unlearning.
- Thorough, *multi-angle evaluation* is essential to detect hidden knowledge or side effects.
- Benchmarks like WMDP, TOFU, and EUD-RELAT help standardize tests but remain incomplete.
- We need an automated *adversarial test suite* plus *holistic metrics* (forget success + retained performance + output quality checks + adversarial recoverability).

---

6. Research Gaps and Future Directions

1. **Truly “Gone for Good” Unlearning**
    - No method fully guarantees knowledge is irrecoverable.
    - The field seeks partial “certified removal” akin to simpler ML models but scaling to LLMs is unsolved.
2. **Robustness to Attacks**
    - Adversarial or paraphrased prompts can re-activate knowledge.
    - Future methods must incorporate adversarial training or additional robust objectives (Zhang et al., 2024; Lynch et al., 2024).
3. **Granular vs. Broad Unlearning**
    - Most tested methods remove entire chunks (e.g., the entire Harry Potter corpus).

    - Entity-level or single-fact unlearning is more challenging, especially if facts are deeply entangled (Ma et al., 2025).
4. **Was Benchmark Expansion**

    - Current standard benchmarks (TOFU, WMDP) are valuable but limited in domain or size.
    - More multi-domain, multilingual sets are needed, plus *continual forgetting* scenarios.
5. **Practical Efficiency**

    - Full-model approaches remain expensive. Parameter-efficient or adapter-based unlearning is an active area but still not widely validated.
    - Real deployments may require repeated unlearning at scale.
6. **Output Quality**

    - Future metrics must ensure the model remains *coherent, factually correct, and helpful* for non-forgotten queries, preventing the "just say I don't know" phenomenon (Yuan et al., 2025).
7. **Towards Standard Evaluation Frameworks**

    - Tools like MUSE, Lynch et al.'s "eight-method suite," and Tofu's metric suite show the community moving toward robust, multi-dimensional testing.
    - Next steps: unify these approaches into a widely accepted "unlearning scoreboard" that tracks forgetting completeness, utility, adversarial re-activation, and output style.

---

7. Implications for Our Project: Building a Robust Evaluation Framework

From surveying the recent literature, it is clear that **evaluation** has become a major bottleneck in advancing unlearning methods for LLMs. The key insights for constructing a *new* or *improved* framework include:

1. **Multiple Angles**:

    - We must measure both *forgetting success* (e.g., accuracy drop on targeted data, membership inference failure) and *utility retention* (accuracy or perplexity on other tasks), as well as *robustness* (adversarial prompts, cross-lingual queries, partial re-training).
2. **Reference "Clean" Model**:

    - Where feasible, compare the unlearned model's outputs on the forget set to a truly *unexposed* or re-trained model's outputs (TOFU's "Truth Ratio" approach). This is the gold standard for verifying if the knowledge is gone, though computationally expensive.
3. **Quality and Semantics**:
    - We should incorporate output-quality checks: token diversity, semantic coherence, factual correctness on unaffected knowledge (Yuan et al., 2025) to catch a model that "solves forgetting" by refusing or giving nonsense answers.
4. **Adversarial Testing**:
    - Evaluate if knowledge can be reactivated by paraphrasing or prompting. Some frameworks systematically try "jailbreak" or re-learning attacks. This is crucial to confirm robust forgetting.
5. **Test Collateral Effects**:
    - Evaluate "neighboring knowledge" or related tasks. If unlearning entity A also partially removes entity B's knowledge, that is an unintended side effect we should measure.
6. **Efficiency**:
    - For real use, measure the computational cost of unlearning, the speed, and memory overhead. Our framework should record these.

By systematically implementing these principles in an automated pipeline with standard benchmarks (TOFU, WMDP, or synthetic sets) and open-sourced metrics, our project can significantly enhance the *comparability and reliability* of LLM unlearning methods.

Conclusion

Machine unlearning for large language models has progressed rapidly in the last year, showcasing diverse fine-tuning, editing, or adapter-based techniques. Yet **evaluation** remains a key obstacle: existing approaches often fail to detect *residual knowledge* or lumps all "incorrect answers" together, missing whether the model is simply refusing queries. Recent literature converges on the need for **multi-dimensional tests**—comparing the unlearned model to a "clean" baseline, measuring adversarial recoverability, checking side effects on normal tasks, and ensuring the model's outputs remain coherent and factual. Benchmarks like **TOFU** and **WMDP** are valuable starting points but must be paired with robust, adversarial checks and behavior metrics.

For our project, the chief takeaway is that **a robust evaluation framework** must:

1. Gather widely used unlearning scenarios (privacy data, harmful knowledge, entity-level forgetting) and benchmarks (TOFU, WMDP).
2. Combine direct forget metrics (accuracy drop) with adversarial tests, utility metrics, and output-quality checks (token diversity, correctness).
3. Include an approach to detect “fake forgetting” via blanket refusals or partial knowledge hiding.
4. Make it easy to scale up to new tasks and repeated unlearning requests.

Such a framework would let the community meaningfully compare unlearning methods, drive real improvements, and avoid deploying techniques that only superficially suppress knowledge. By consolidating these insights and evaluation best practices, we can push the field toward *truly effective and verifiable* machine unlearning in LLMs.

References (cited in text)

- Chen et al. 2024, *LoRA-based Unlearning.*
- Chen & Yang 2023, *“EUL: Efficient Unlearning Layers.”*
- Doshi & Stickland 2024, *“Does Unlearning Truly Unlearn?”*
- Eldan & Russinovich 2023, *“Who’s Harry Potter? Approximate Unlearning in LLMs.”*
- Guo et al. 2024, *Mechanistic Unlearning.*
- Kim et al. 2024, *NegMerge: Weight Negation.*
- Li et al. 2024, *RMU and WMDP Benchmark.*
- Liu et al. 2024, *“SKU: Towards Safer LLMs through Machine Unlearning.”*
- Lynch et al. 2024, *“Eight Methods to Evaluate Robust Unlearning in LLMs.”*
- Ma et al. 2025, *Entity-Level Unlearning.*
- Maini et al. 2024, *TOFU (Task of Fictitious Unlearning)*.
- Pan et al. 2024, *Multi-Objective LLM Unlearning.*
- Wu et al. 2024, *Synthetic Knowledge Bases for Deep Unlearning.*
- Yao et al. 2023, *LLMU: Negative Gradient Fine-tuning.*
- Yuan et al. 2025, *Adding Output-Quality Metrics (Token Diversity, Coherence, Factuality).*
- Zhang et al. 2024, *Quantization Attacks & Partial Recovery of Forgotten Data.*

(References to “2410.14425” or “2412.20412” align with preprint numbering in the conversation text, indicating the year+arXiv ID.)

End of Literature Review

⛔ Expectations and Failure modes

Additionally, if you'd like, you can also go through this doc, by Michael Aird: *[shared] Research Project Planning Template* *. Diogo will write his opinion after everyone else, so as to not influence what you write.*

Name	**Expectations** What you expect to get out of the project	**Failure modes**		
		Problem we may run into	**Impact** of problem on project	**Mitigation** for the problem
Diogo Cruz	-	-	-	-
Alexander Panfilov				
Ashwin Sreevatsa	Learn a lot about the current state of unlearning Get hands on experience with developing an unlearning pipeline setup Hands-on experience developing an unlearning benchmark?	• People get slotted into roles in the project that they might not be as excited about • Due to remote work, people end up very siloed (working at different times, etc)	• Less enthusiasm for the project, slower progress as a result • Makes collaboration harder, people might redo similar work or be unaware of work others are doing that might be relevant for them	• regular check-ins about what is going well, not well might help. • Establishing coworking times, sync vs async communication expectations
Jan Batzner	(1) get on top of Unlearning Literature; (2) gain remote-only project experience; (3) explore the benchmarking problem; (4) learn from teammates with stronger software engineering skills	Research question not clear enough; similar papers being published simultaneously; compute budget limited; getting started time takes over majority of project duration	Important to keep aiming for MVP; all of these problems manageable when communicated and discussed	Regular check-ins and state research problem and our hypotheses, be up to date on related works, use university GPU if we run out of compute, start early drafting a paper to guide discussions
Shariqah Hossain	- Create a robust evaluation framework for unlearning - Learn about and tackle challenges of short-term research projects, as well as challenges unique to AI safety and evaluations - Gain experience in deciding research directions and producing results in a reliable, well-communicated way - Collaborate with others interested in AI safety	Evaluating on a variety of data that isn't use-case specific like existing benchmarks	There isn't a state-of-the-art benchmark to go off of, it may take longer to curate the datasets we need for evaluation	- Start small on existing datasets - Focus on creating a generalized benchmark - Look at benchmarks for other non-unlearning use cases as a reference for topics to include

Yeonwoo Jang	Gain practical experience in building eval suites, open sourcing an eval benchmark and code (and a blog/paper?!) that help track progress in machine unlearning	coming up with good metrics — current metrics/eval frameworks can be misleading as models may appear to forget but retain recoverable knowledge; time/resource constraints	could lead to false confidence in unlearning methods and waste research effort, technical limitations might slow development and miss critical failure modes	start with smaller models/datasets for rapid development and include adversarial testing from the beginning; implement multiple complementary metrics including robustness checks; use modular design and efficient eval strategies (batching, caching) to handle resource constraints

Original proposal

This is a private copy of the proposal, that we can modify as needed. The original can be found in Project Proposal: Robust Evaluation Framework for LLM Unlearning.

Robust Evaluation Framework for LLM Unlearning

Summary

This project aims to develop a practical evaluation framework for machine unlearning in Large Language Models (LLMs). Building on recent work in unlearning methods (Eldan & Russinovich, 2023; Yao et al., 2023; Li et al., 2024; Liu et al., 2024), we will focus on creating standardized metrics, benchmark datasets, and evaluation tools to assess both unlearning completeness and model capability preservation. Our research will help bridge the gap between theoretical unlearning approaches and their practical effectiveness, contributing to more reliable and verifiable unlearning techniques.

Motivation and Background

Recent advances in machine unlearning for LLMs have shown promising results in removing specific information from models. However, current evaluation methods have several significant limitations. Simple prompt-based testing often misses residual knowledge, and there's no standardized way to measure the trade-off between unlearning and capability preservation. Additionally, unlearned information can sometimes be recovered through careful prompting, and existing metrics don't account for consistency across different contexts.

These limitations create substantial challenges for the field. It becomes difficult to compare different unlearning approaches reliably or ensure that unlearning methods are truly effective rather than just masking information. Understanding the limitations and failure modes of current techniques also remains a significant challenge.
The project will consist of five main components:

1. Develop Core Evaluation Metrics
2. Create Initial Benchmark Datasets
3. Build Testing Pipeline
4. Conduct Evaluation Study
5. Document Findings

For the evaluation metrics, we'll design measures for unlearning completeness while creating ways to assess model capability preservation and implement consistency measures across different contexts. Our benchmark datasets will include test cases for common unlearning

scenarios, with systematic variations of test queries and control tasks for measuring general capabilities.

The testing pipeline will feature an automated evaluation framework with standardized reporting formats and basic visualization tools. We'll then use this infrastructure to test existing unlearning methods, analyze basic failure modes, and compare effectiveness across approaches.

Throughout the project, we'll maintain clear documentation for the framework, prepare our results for publication, and create practical guidelines for practitioners.

Potential Challenges and Backup Plans

We've identified several key challenges and developed backup plans for each. First, we may face limited compute resources for testing. In this case, we'll focus on smaller models and fewer but more carefully chosen test cases. Second, creating comprehensive benchmarks in our limited time could prove difficult. Our backup plan is to start with a smaller, well-defined set of test cases and design for easy extensibility. Third, measuring capability preservation can be complex. We'll begin with basic capability metrics and gradually expand based on our findings.

Scope and Ambition

Our least ambitious version would develop basic evaluation metrics, create a small benchmark dataset, build a simple testing pipeline, test on 1-2 open-source models, and release an initial toolkit. In our most ambitious version, we would develop a comprehensive evaluation framework with extensive benchmark datasets, a sophisticated testing pipeline, tests across multiple models and methods, and publication-ready results.

Output

We aim to produce an open-source evaluation framework, initial benchmark datasets, documentation and usage guidelines, a research paper draft, and a blog post summarizing our findings.

Theory of Change

Our project contributes to AI safety through several key pathways. For direct impact, we enable reliable testing of unlearning methods and provide tools for researchers to identify weaknesses in current approaches, helping prevent deployment of ineffective methods that could give a false sense of security.

In terms of field building, we're creating standardized evaluation methods that enable better research while making it easier to compare different approaches and identify promising directions. This helps build a shared understanding of what constitutes effective unlearning.

Looking toward future impact, we're setting a foundation for evaluating more sophisticated unlearning methods as models become more capable. Our work helps identify fundamental limitations in current approaches and creates tools that can evolve alongside advancing AI capabilities.

For knowledge distribution, our open-source tools make rigorous evaluation accessible to more researchers, while our documentation and guidelines help spread best practices. Our published findings will contribute to the shared knowledge base of the field.

This work matters because as AI systems become more capable, having reliable ways to remove information becomes crucial for safety. The current lack of standardized evaluation makes it hard to develop better methods, and without robust testing, we risk deploying ineffective solutions that could fail when most needed. By improving how we evaluate unlearning methods now, we help ensure that future work in this area builds on solid foundations.

Team Structure

Team Size: 2-4 mentees

Mentor

Diogo Cruz (email: diogo.abc.cruz@gmail.com)

I have experience in AI safety research, having led a team in AI Safety Hub Labs (now LASR) that resulted in a NeurIPS SoLaR paper (https://arxiv.org/abs/2311.04046). I have also worked on a project analyzing learned look-ahead behavior in chess neural networks. I have a lot of research experience from my Quantum Computing PhD (Google Scholar), during which I also mentored Bachelor and Masters' students.

This project builds directly on my work replicating the WMDP unlearning benchmark (see https://github.com/jeanne-s/wmdp_rec and Research Engineers Guide.

For this project, I will either co-work or provide detailed guidance, depending on the team's needs.

Required Skills:

- Python programming
- Basic ML/NLP knowledge
- Familiarity with transformers/LLMs
- Good scientific writing

Nice to Have:

- Experience with evaluation metrics
- Knowledge of machine unlearning
- Familiarity with ML testing frameworks

Roles: While all team members will contribute across areas, potential focus areas include:

- Metrics development
- Benchmark creation
- Pipeline implementation
- Experimental analysis

Timeline

(This is tentative, and heavily dependent on the team's skill set and project pace)
The project will span three months:

Month 1: We'll focus on project setup, background research, initial metric design, and basic benchmark creation.

Month 2: Our efforts will center on framework implementation, pipeline development, and preliminary testing.

Month 3: We'll conduct our full evaluation study, complete our analysis and documentation, and focus on paper/blog writing.

Weekly meetings and regular check-ins will help maintain progress and address challenges as they arise.

ℹ️ About

This is essentially a summary of what Diogo mentioned to most team members during the interview.

My background

I have a bachelor's and master's in physics and a PhD in quantum computing. Around halfway through my PhD, before ChatGPT was released, I became increasingly interested in AI safety and started doing part-time projects in this space. I've worked on projects related to:

- Inductive biases in large language models
- Mechanistic interpretability
- Evaluations (my current focus)

For the past ~6+ months, I've been doing AI safety research full-time, primarily focused on evaluations. My current projects include:

- Evaluations for autonomous agents
- Multi-turn jailbreaks
- Machine unlearning (this SPAR project)

Purpose

This project aims to develop a robust evaluation framework for assessing machine unlearning techniques in large language models. While various unlearning methods exist to remove specific information or capabilities from models, current evaluation approaches often fail to comprehensively test whether the information has been truly unlearned versus just suppressed in certain contexts. We will work to create standardized benchmarks and evaluation methods that can test unlearning claims from multiple angles to enable more reliable assessment of unlearning techniques.

Planned type and degree of involvement

I plan to be quite hands-on early in the project and then gradually shift to a more supervisory role as the team develops momentum. My involvement will include:

- One mandatory 1-hour weekly team meeting for sync-ups and updates
- Optional co-working sessions using Gather Town for collaborative work
- Communication via SPAR Slack
- Code collaboration through GitHub
- Documentation in Google Docs initially, transitioning to Overleaf for paper writing
- Particularly active involvement in the first few weeks to help establish direction and momentum

What I'm asking for

- Time commitment: Hopefully at least 10 hours per week. 5h or less would mean that you would spend most of your time just catching up to the rest of the team.
- Strong preference for being available during the first few weeks to build initial momentum
- Clear communication about any scheduling conflicts or availability changes
- Willingness to coordinate across time zones (team split between US and European time zones)

Research question and sub-questions

Core questions include:

- How can we create comprehensive benchmarks that test unlearning claims from multiple angles?
- What are the gaps in current evaluation approaches for machine unlearning?
- How can we test whether information is truly unlearned versus just suppressed in certain contexts?
- What role can model internals analysis play in verifying unlearning claims?

Scope

MVP (Goal 1):

- Collect and unify existing evaluation techniques from literature into a standardized benchmark framework
- Create initial implementation that others can use to test their unlearning approaches

Stretch goals (Goals 2-3):

- Extend benchmarks to cover multiple languages, formats (code, math notation, etc.)
- Develop automated approaches for generating robust test cases
- Explore model internals analysis approaches (e.g., using sparse autoencoders)
- Adversarial testing to find gaps in evaluation coverage

Out of scope:

- Developing new unlearning techniques
- Full analysis of all possible evaluation angles
- Extremely compute-intensive approaches

Methods

The project will involve:

- Literature review (especially first week) to understand current approaches
- Engineering work to implement evaluation framework
- Iterative testing and refinement of benchmarks
- Possible compute budget of $500 for running evaluations

- Regular sync-ups to coordinate parallel workstreams
- Documentation and paper writing

Output

Expected outputs:

- Most likely: Workshop paper (5 pages + appendices) that we can submit to an ML conference workshop
- Minimum: Blog post sharing results if project faces major obstacles
- Stretch goal: Full conference paper (8-10 pages) if project goes particularly well
- Public code repository with evaluation framework
- Possible benchmarks published on Hugging Face

The project timeframe is approximately 3 months (mid-February to mid-May 2025). All meaningful contributors will be included as co-authors on any publications.

Tab 9

Checklist for final submission

- Repo
    - Make it paper standard
    - What are the components we need?
        - Table 1, 2 ✅
            - inference.py
        - Table 3 ✅
            - Lm-eval inference on 8 models ✅
            - Custom tasks ✅
            - Rephrased data (not provided but the prompts are provided in the paper Appendix) ✅
        - 5-shot stuff (##TODO)
    - Probe
        - Move this out of the plots/ folder ✅
    - Plot
        - ~~Add my code~~
        - Not really needed?
- Paper
    - COLM carema-ready version
    - ArXiv version
- Wins