

CS 839: Foundation Models Security, Privacy, Toxicity Presentation & Project Info

Fred Sala

University of Wisconsin-Madison

Nov. 2, 2023

Announcements

•Logistics:

- HW2 due Nov. 9. One more (small) homework after
- Project info out. Dates: Nov. 16: proposal, Dec. 19
- Presentation dates: **Nov:** 9,14,16,21,28,30 **Dec:** 5,7
 - Two slots/date (worst-case three might be needed in some special cases)
- Due On Nov. 2 (Today!) --- presentation!

•Class roadmap: (note change!)

Thursday Nov. 2	Presentation + Project Info
Tuesday Nov. 7	Security, Privacy, Toxicity + The Future

Presentation Hints: For Presenters

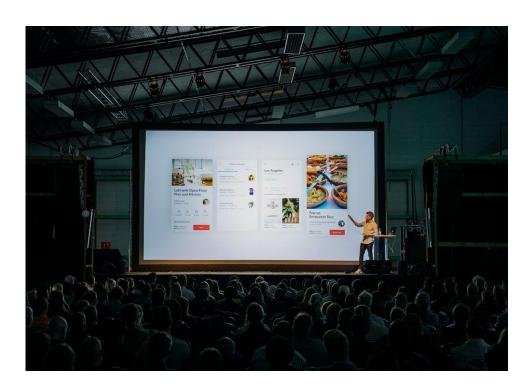
- •Why did you select this paper?
- How does it fit into our class?
 - Example: studies PEFT variations
- •What does it do beyond what other papers?
- •What are the implications?
 - Who will care?
- •What are insights beyond paper?
 - From you!



Presentation Hints: For Presenters

- •What are insights beyond paper?
 - From you!
- •This is the bit we're most interested in.
- Can describe flaws or limitations you've found
- Or new applications
- Or extensions
- Or inspiration you've gotten

Please have a few slides on this!



Presentation Hints: For Presenters

- Finally, we want to have audience engagement
- Please take a minute to ask audience (and allow volunteers to answer) 1-2 questions
 - To gauge understanding
 - Or to drive a discussion forward



Presentation Hints: For Audience

Please engage & ask questions!

- We won't formally track this, but we are hoping for a conversation/discussion
- Obviously please attend.
- Rough goal: each of you please ask two questions!
 - Clarifying questions are fine

Project Hints

Your project should have the following:

- Hypothesis (clearly state it!)
- Some means of verifying it

- Contextualization for the work
- Related work
- Future work



Project Hints: Writing

Very generally, you can structure your report as follows:

- Also great for papers in general,
- Ignore parts that are not directly relevant

Tips for Writing Technical Papers

Jennifer Widom, January 2006

Here are the notes from a presentation I gave at the <u>Stanford InfoLab</u> Friday lunch, 1/27/06, with a few (not many) 10 revisions for the 10/19/12 revival. The presentation covered:

- Paper Title
- The Abstract
- The Introduction
- · Related Work
- The Body
- Performance Experiments
- The Conclusions
- Future Work
- The Acknowledgements
- Citations
- Appendices
- Grammar and Small-Scale Presentation Issues
- Mechanics
- · Versions and Distribution

https://cs.stanford.edu/people/widom/paper-writing.html

Project Hints: Writing an Intro

Intro can be challenging to write

- For your project report,
- Also for a paper
- Template is pretty good!
 - Answer each question with a bullet outline, then write
- 1. What is the problem?
- 2. Why is it interesting and important?
- 3. Why is it hard? (E.g., why do naive approaches fail?)
- 4. Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions? How does mine differ?)
- 5. What are the key components of my approach and results? Also include any specific limitations.

Project Hints: Writing Experiments

Often challenging: structuring an experiment section!

- Hypothesis
- Proxy
- Protocol
- Expected Results
- Results

How to setup an experiment

The Coffee Experiment!



https://aberger.site/coffee

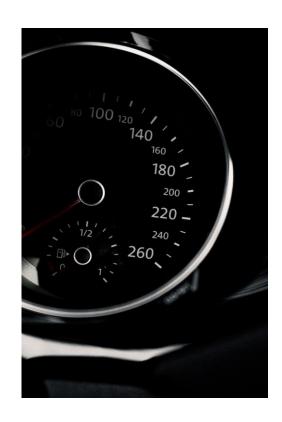
Writing: Paper Hints

Beyond this class, one trick to keep in mind.

- Most reviewers not as knowledgeable as you are in your area
- Do not have access to a simple metric to evaluate your work

•Provide it to them!

- Otherwise, they will default to heuristics
 - Do you achieve SOTA?
 - Is the cosmetic writing/appearance good?
 - Other often irrelevant things





Thank You!