Curriculum

Mini-Curriculum for *Ctrl-Z* and the Wider AI-Safety Landscape

A six-report roadmap you can develop alongside your fortnightly journal club.

Report 0 (½-page cheat-sheet) – "Where does Ctrl-Z sit?"

Axis	Position of the paper
Threat model	*Subversive* or **untrusted** LLM agents that may carry out covert attacks (e.g. download malware).
Risk-mitigation strategy	**Control evaluations** – ex-ante safety layers (monitoring, resampling, auditing) that keep a system safe *even if* the core model is malicious.
Novelty	First control evaluation in a **multi-step agent** setting; introduces **BashBench** and **resample protocols** that cut attack success from 58 % to 7 % at only 5 % usefulness cost.
Lineage	Extends Greenblatt et al. 2023 (single-step control evaluations); contrasts with Wen et al. 2024 (distributed-threat setting).

Report 1 – *AI-Safety Field Guide* (the "big picture")

Goal: build a high-altitude map (one diagram + ~4 pages) showing the main sub-fields:

1. **Preventive Alignment** – RLHF, constitutional AI, interpretability.
2. **Control & Oversight** – debate, scalable oversight, **Ctrl-Z** lineage.
3. **Robustness & Adversarial ML** – adversarial attacks, data poisoning.
4. **Emergent-risk Studies** – deception, goal mis-generalisation.
5. **Governance, Audits & Safety Cases** – standards, policy, ISO-42001 parallels.

Indicate where *Ctrl-Z* lives ("Control & Oversight") and draw arrows to adjacent boxes.

Report 2 – *Control Evaluations & Red-Team Methodology*

1. **Single-step control** – original APPS back-door evaluations.
2. **Multi-step complications** – why long-horizon tasks enable "low-per-step" attacks.
3. **BashBench dataset** – how sys-admin tasks were curated.
4. **Protocol design space**
   - Trusted vs untrusted monitoring.
   - **Defer-to-resample** vs defer-to-trusted.
5. **Metrics** – usefulness ↔ safety Pareto; cost–safety trade-off curves.
6. **Open questions** – longer trajectories, imperfect audits, strategic attack selection.

Parallel reads: Wen 2024 (distributed threat), Mallen 2024 (subversion strategies).

Report 3 – *Deception & "Fake Alignment"*

- Carlsmith 2023 "Scheming AIs".
- Greenblatt 2024 "Alignment-Faking in LLMs".
- Mechanistic case studies of deceptive circuits.
- How control protocols (e.g., resampling) interact with deceptive behaviour.

Report 4 – *Robustness & Adversarial ML for LLM Agents*

- Data-poisoning & back-door attacks.
- Evasion attacks vs control defences.
- Universal jailbreaks as input-level analogues.
- How robustness methods complement control evaluations.

Report 5 – *Scalable Oversight & Weak-to-Strong Generalisation*

- Bowman 2022 on measuring scalable oversight progress.
- Burns 2023 "Weak-to-Strong generalisation".
- Debate, recursive reward modelling, RLAIF.
- How **Ctrl-Z** uses a weaker monitor model + human audits.

Report 6 – *Governance & Safety-Case Frameworks*

- Safety cases for advanced AI (Clymer 2024).
- ISO/IEC 42001 safety-management analogies.

- Policy implications: audit requirements, deployment licences, red-team transparency.