PPRGS Framework - Setup Checklist

Complete these steps to get PPRGS ready for publication and testing.

Day 1: Repository Setup

GitHub Repository Creation

- [ ] Create new GitHub repo: `PPRGS-Framework`
- [ ] Make it public
- [ ] Add description: "Making wisdom the goal: An AI alignment framework for ASI adaptability"
- [ ] Add topics: `ai-safety`, `alignment`, `asi`, `reinforcement-learning`, `artificial-intelligence`

File Structure Setup

- [ ] Run `./setup.sh` to create directory structure
- [ ] Copy all artifact files to appropriate locations:
- [ ] `README.md` → root
- [ ] `LICENSE` → root
- [ ] `CODE_OF_CONDUCT.md` → root
- [ ] `CONTRIBUTING.md` → root
- [ ] `CONTRIBUTORS.md` → root
- [ ] `requirements.txt` → root
- [ ] `.gitignore` → root
- [ ] `QUICKSTART.md` → docs/
- [ ] `FAQ.md` → docs/
- [ ] `rv_calculator.py` → metrics/
- [ ] `ees_fduds.py` → metrics/
- [ ] `pprgs_agent.py` → implementations/gpt4/
- [ ] `run_test.py` → experiments/experiment_2_enrichment/
- [ ] `pprgs_arxiv.tex` → paper/

Initial Commit

- [ ] `git init`
- [ ] `git add .`
- [ ] `git commit -m "Initial commit: PPRGS Framework v1.0"`
- [ ] `git branch -M main`
- [ ] `git remote add origin git@github.com:YOUR_USERNAME/PPRGS-Framework.git`
- [ ] `git push -u origin main`

GitHub Settings

- [ ] Enable Issues
- [ ] Enable Discussions
- [ ] Add repository description
- [ ] Add website (your LinkedIn): https://www.linkedin.com/in/michael-riccardi-b3b41695/
- [ ] Create initial labels:
  - `bug`, `enhancement`, `documentation`, `good first issue`
  - `experiment-replication`, `red-team`, `help-wanted`
  - `security`, `platform-aws`, `platform-gpt4`, `platform-gemini`, `platform-grok`

Day 2: arXiv Submission

Paper Preparation

- [ ] Copy full paper content into `paper/pprgs_arxiv.tex`
- [ ] Compile locally: `pdflatex pprgs_arxiv.tex`
- [ ] Fix any LaTeX errors
- [ ] Generate bibliography: `bibtex pprgs_arxiv`
- [ ] Re-compile twice more for references
- [ ] Check PDF output

arXiv Account

- [ ] Create account at arxiv.org (if needed)
- [ ] Complete user profile
- [ ] Get endorsement if required (cs.AI usually doesn't need it)

Submission

- [ ] Submit to arxiv.org
- [ ] Primary category: `cs.AI` (Artificial Intelligence)
- [ ] Cross-list: `cs.CY` (Computers and Society)
- [ ] Cross-list: `cs.LG` (Machine Learning)
- [ ] Upload compiled PDF
- [ ] Double-check all metadata (title, author, abstract)
- [ ] Submit!
- [ ] **Save arXiv ID** (format: 2025.XXXXX)

Post-Submission

- [ ] Update README.md with arXiv badge and link

- [ ] Update paper citation in all docs with arXiv ID
- [ ] Tweet/post announcement with arXiv link

Day 3: Code Testing

Environment Setup

- [ ] Create `.env` file with your OpenAI API key
- [ ] Test virtual environment: `source venv/bin/activate`
- [ ] Verify all packages installed: `pip list`

Core Metrics Testing

- [ ] Run `python metrics/rv_calculator.py`
- [ ] Verify demonstration output looks correct
- [ ] Run `python metrics/ees_fduds.py`
- [ ] Verify RC enforcement logic works

GPT-4 Agent Testing

- [ ] Run `python implementations/gpt4/pprgs_agent.py`
- [ ] Verify 4 function calls on MRP (task 3)
- [ ] Check R_V calculation makes sense
- [ ] Verify metrics export works

Experiment 2 Testing

- [ ] Run baseline: `python experiments/experiment_2_enrichment/run_test.py --agent ums --trials 3`
- [ ] Run PPRGS: `python experiments/experiment_2_enrichment/run_test.py --agent pprgs --trials 3`
- [ ] Run comparison: `python experiments/experiment_2_enrichment/run_test.py --agent both --trials 3`
- [ ] Verify success criteria are checked correctly
- [ ] Check results JSON export

Bug Fixes

- [ ] Document any bugs found
- [ ] Fix critical issues
- [ ] Create GitHub issues for non-critical bugs
- [ ] Commit fixes: `git commit -m "Fix: [description]"`
- [ ] Push: `git push`

Day 4: Community Outreach

Alignment Forum

- [ ] Create account (if needed)
- [ ] Write post summarizing PPRGS
- [ ] Include arXiv link
- [ ] Include GitHub link
- [ ] Post in AI Alignment category
- [ ] Respond to comments

LessWrong

- [ ] Cross-post or link to Alignment Forum post
- [ ] Engage in technical discussions
- [ ] Address counterarguments constructively

Twitter/X

- [ ] Write thread (8-12 tweets) covering:
  - Problem (over-optimization paradox)
  - Solution (PPRGS framework)
  - Key innovation (R_V formula with multiplication)
  - Results (87% higher R_V in Exp 2)
  - Call to action (replicate experiments)
- [ ] Include arXiv link
- [ ] Include GitHub link
- [ ] Tag relevant researchers/organizations
- [ ] Use hashtags: #AISafety #AIAlignment #MachineLearning

LinkedIn

- [ ] Write professional post about PPRGS
- [ ] Target AI safety professionals
- [ ] Link to arXiv and GitHub
- [ ] Mention it's open for collaboration

Email Outreach (Optional for Day 4, can wait until Day 5-7)

- [ ] Draft email to key researchers (use template from publication plan)
- [ ] Send to 3-5 researchers for initial feedback

Day 5-7: Iteration & Improvement

Community Engagement

- [ ] Respond to all GitHub issues within 24 hours
- [ ] Answer questions on Alignment Forum/LessWrong
- [ ] Engage on Twitter
- [ ] Update FAQ based on common questions

Documentation Improvements

- [ ] Add any missing clarifications to README
- [ ] Create diagrams if people request them
- [ ] Add example outputs/screenshots
- [ ] Improve error messages based on user reports

Code Refinements

- [ ] Fix bugs reported by early testers
- [ ] Add better error handling
- [ ] Improve logging/debugging options
- [ ] Add progress bars for long-running experiments

First Replication Attempt

- [ ] Help first person trying to replicate
- [ ] Document their process
- [ ] Fix any blockers they encounter
- [ ] Celebrate their success publicly!

Week 2: Platform Expansion

AWS Bedrock Implementation (Optional)

- [ ] Complete CloudFormation templates
- [ ] Test deployment
- [ ] Document setup process
- [ ] Create tutorial

Gemini Implementation (Optional)

- [ ] Complete multimodal P_2 implementation
- [ ] Test with images/video
- [ ] Document results

Experiment 1 & 4 (Optional)

- [ ] Build remaining experiment environments
- [ ] Run initial tests
- [ ] Document results

Metrics & Milestones

Week 1 Goals

- [ ] arXiv paper published: Yes/No
- [ ] GitHub stars: Target 50+
- [ ] Experiment 2 replications: Target 1+
- [ ] Community posts: Target 3+ platforms

Week 2 Goals

- [ ] GitHub stars: Target 100+
- [ ] Independent replications: Target 3+
- [ ] Pull requests: Target 2+
- [ ] Media mentions: Target 1+

Month 1 Goals

- [ ] All 4 experiments validated
- [ ] Cross-platform validation started
- [ ] Conference submission: Yes/No
- [ ] Active contributors: Target 5+

Contact Information Updates

Update these with your actual handles once created:

- [ ] GitHub username in all docs
- [ ] Discord server link (if you create one)
- [ ] Twitter handle (if you create dedicated account)
- [ ] Update all placeholder URLs

Critical Reminders

Before First Push

- [ ] Remove any API keys from code

- [ ] Check `.gitignore` is working
- [ ] Verify no sensitive data in commits
- [ ] Double-check LICENSE is correct

Before arXiv Submission

- [ ] Spell-check entire paper
- [ ] Verify all citations are correct
- [ ] Check all equations render properly
- [ ] Read abstract out loud (catch errors)

Before Experiment Runs

- [ ] Set API rate limits appropriately
- [ ] Monitor costs (especially on AWS)
- [ ] Have backup of results
- [ ] Document all hyperparameters

Emergency Contacts

If something goes wrong:

- **GitHub issues**: For code problems
- **arXiv help**: help@arxiv.org
- **OpenAI support**: For API issues
- **Community**: Alignment Forum for conceptual questions

Progress Tracker

Current Phase: [ ] Day 1 | [ ] Day 2 | [ ] Day 3 | [ ] Day 4-7 | [ ] Week 2+

Blockers: (List any issues preventing progress)

Wins: (Celebrate successes!)

Next Action: (What's the very next thing to do?)

Remember: Ship fast, iterate publicly, engage with feedback. The window is closing—let's make this count!

📧 Questions? mike@mikericcardi.com