

7.4 Appendix B: Experiment Style Checklist

HOW to organize and present artifacts for easier reviewer use

Structure

- Is the experiment organized into distinct phases (setup, execution, analysis)?
- Is a top-level guide through the complete workflow provided?
- Are components modularized to allow partial execution, updates, and testing?
- Is a logical directory structure with descriptive names created?

Validation

- Are verification steps included after each major phase?
- Are expected output examples provided for validation points?
- Are automated tests that verify correct execution included?
- Are known warnings that can be safely ignored documented?
- Is troubleshooting guidance for common issues offered?

Configuration

- Is there a single configuration file or mechanism that controls all experiment parameters?
- Are configuration files or environment variables used instead of hardcoding values (including logins and passwords for accessing APIs/services)?
- Are all changeable values (paths, hostnames, file locations) parameterized?
- Are reasonable defaults included for all configurable parameters?
- Is experiment logic separated from environment-specific settings?
- Are there sanity checks that confirm the experimental environment is properly configured?

Reproducibility Condition

- Is a clear definition provided for what constitutes successful reproduction? Are both exact reproducibility criteria (e.g., bit-for-bit output matching) and functional reproducibility criteria (e.g., statistically equivalent results) defined (if applicable)?
- Are acceptable thresholds for numerical result variations specified?
- Are quantitative metrics provided to evaluate reproduction success?
- Are known sources of non-determinism and their expected impact documented? Are statistical methods provided to evaluate reproducibility (confidence intervals, error bars)?
- Are there tools or scripts to automatically compare and validate outputs against expected results?

Time & Resource Estimation

- Are overall runtime estimates for the complete experiment provided?
- Is there a breakdown of time estimates for each major phase? Do distinct experiment stages have clear indicators of progress (i.e., a progress bar or estimated time left)?
- Are points where significant waiting periods occur indicated?
- Are any background processes that may impact timing noted?
- Are resource scaling properties documented (e.g., how requirements change with input size)?
- Is there guidance on optimizing resource usage for different hardware environments?