

## PromptSnap: Community Standards and Datasets for Natural Language Programming Assessment

**Overview:** GenAI programming assistants have transformed programming into a dialogue between natural language and code, elevating two competencies: *code comprehension* (understanding and explaining code) and *prompt engineering* (articulating programming intent to AI). Two assessment types target these skills: “*Explain in Plain Language*” (EiPL) questions [2, 4] present students with code and ask them to describe what it does in natural language (e.g., “Explain what this function accomplishes”), assessing students ability to accurately and abstractly describe the purpose of code. *Prompt problems* [1] invert this with students receiving a specification (typically input-output examples showing desired behavior) and must write a natural language prompt that would cause an AI assistant to generate code exhibiting that behavior.

Despite promise for evaluating GenAI-era programming skills, the field lacks standardized data formats and shared datasets with each tool appearing to have independent formats for responses, rubrics, and metadata, which increases the difficulty of conducting cross-institutional research, comparisons, and replication. This proposal aims to create *PromptSnap*, a standardized data format and tooling for natural language programming assessment, complementing—or extending—SPLICE’s ProgSnap2 [3], which captures programming process data.

### Deliverables

(1) CSV-based data format specification, (2) open-source tools for multi-institutional deployment which adhere to that format, (3) initial datasets from pilot tests to be hosted on DataShop and the SPLICE Dataset Catalog.

**PromptSnap Specification and Tooling Infrastructure:** PromptSnap aims to capture natural language programming assessment data, specifically student responses to EiPL questions (explaining code) and Prompt Problems (crafting prompts to generate code), evaluating GenAI-era comprehension and communication skills. Following ProgSnap2’s model, PromptSnap will aim to maintain field compatibility in core structures (e.g., SubjectID, CourseID, ...), using the same CSV-based format with MainTable.csv and DatasetMetadata.csv. PromptSnap aims to extend ProgSnap2 by adding fields for natural language responses, grading outcomes where LLMs generate code from student responses and test functional correctness, comprehension level classifications, reference solutions with unit tests, and natural language programming specific revision patterns.

Leveraging PI Smith’s existing collaborations with tool developers at the University of Auckland (who developed Prompt Problems and the Promptly platform), practitioners at UCSD, Auckland, and UIUC (who have deployed EiPL questions), and the SPLICE LLM working group, we will collaboratively develop this standard through an iterative community feedback process. This design ensures the robustness to support diverse natural language assessment approaches while maintaining compatibility with ProgSnap2’s.

Open-Source Tooling: This project will scale prototype tools developed by PI Smith to production-ready, multi-institutional deployment: (1) *eiplgrader*<sup>1</sup>—an open-source autograding library implementing code generation-based grading (using LLMs to generate code from student explanations, then testing functional correctness) and prompt segmentation (dividing responses into semantic units for comprehension assessment), (2) *Purplex*<sup>2</sup>—a web-based platform for deploying EiPL questions and other natural language programming activities with automated feedback, and (3) *PrairieLearn integration*<sup>3</sup>—enabling PromptSnap-compliant question deployment and data collec-

<sup>1</sup><https://github.com/CoffeePoweredComputers/eiplgrader>

<sup>2</sup><https://purplex.org>

<sup>3</sup><https://us.prairielearn.com/>

tion to the 180k+ students at 20+ institutions currently served by PrairieLearn. These tools will support both EiPL questions and prompt problems, with extensible architecture for community contributions.

**Dataset Goals:** Upon successful tool deployment and IRB approval, we will collect and curate initial datasets exemplifying the PromptSnap specification. These datasets—including student responses, grading outcomes, and rich metadata—would be contributed to DataShop and the SPLICE Catalog, providing the community with concrete examples and benchmarks for natural language programming assessment research.

**Community Impact:** PromptSnap aims to develop standardized data formats for emerging GenAI focused tools and activities, enabling reproducible cross-institutional research. By providing common infrastructure, PromptSnap will enable researchers to: compare autograding approaches using common benchmarks, conduct cross-institutional studies of prompting and comprehension skill development, train and evaluate learner models, and reduce barriers to studying natural language programming assessment. Educators gain access to empirically validated question banks, reusable open-source tools, and automated grading systems deployable without programming expertise—lowering adoption barriers for these critical GenAI-era assessments.

**Project Team:** *David H. Smith IV* (PI, Virginia Tech) developed EiPL autograding approaches, the *eiplgrader* package, and the Purplex platform. His work focuses on developing and evaluating AI-integrated learning tools for computing education, with collaborations at universities in the U.S., New Zealand, and India. *Undergraduate researchers* guided by PI Smith and graduate students will support tool development, support multi-institutional deployments, and curate collected data under PI supervision.

**12-Month Timeline and Budget:** *Months 1–9:* Initial specification development through community feedback; *Months 4–9:* Tool development, multi-institutional deployment coordination; *Months 10–11:* Data collection and curation; *Month 12:* Documentation, dataset release, and dissemination.

| Item               | Purpose                                           | Amount          |
|--------------------|---------------------------------------------------|-----------------|
| Student wages      | Tool development & deployment (206 hrs @ \$25/hr) | \$5,142         |
| Fringe benefits    | Student benefits (7.28%/6.73%)                    | \$351           |
| Travel             | Collaboration visits & conference presentation    | \$1,500         |
| Computing services | Platform hosting & GenAI API credits              | \$5,000         |
| Indirect costs     | University overhead (61% MTDC)                    | \$7,316         |
| <b>Total</b>       |                                                   | <b>\$19,309</b> |

## References

- [1] P. Denny, J. Leinonen, J. Prather, A. Luxton-Reilly, T. Amarouche, B. A. Becker, and B. N. Reeves. Prompt problems: A new programming exercise for the generative ai era. SIGCSE '24.
- [2] L. Murphy, R. McCauley, and S. Fitzgerald. 'explain in plain english' questions: Implications for teaching. SIGCSE '12.
- [3] T. W. Price, D. Hovemeyer, K. Rivers, G. Gao, A. C. Bart, A. M. Kazerouni, B. A. Becker, A. Petersen, L. Gusukuma, S. H. Edwards, and D. S. Babcock. Progsnap2: A flexible format for programming process data. ITiCSE '20.
- [4] D. H. Smith IV and C. Zilles. Code generation based grading: Evaluating an autograding mechanism for "explain-in-plain-english" questions. ITiCSE '24.