2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50804

WIP: Strategies and Practices to Overcome Irresponsible and Low-Effort Deliverables (SPOILED) in an Excel-Based Database Class

Dr. Win P. V. Nguyen, Virginia Polytechnic Institute and State University

Dr. Win Nguyen is a Collegiate Assistant Professor in the Grado Department of Industrial and Systems Engineering at Virginia Tech. He earned his Ph.D. in Industrial Engineering from Purdue University. His primary focus and passion is undergraduate engineering education, where he is dedicated to fostering student growth and development, whether a student in his classes, is a teaching assistant or research assistant, or a senior design student. He is also committed to coaching students on study skills, technical problem-solving, career preparation, and professional development. His research interests include applying undergraduate data analytics to enhance student learning and career outcomes, as well as the responsible and effective use of generative AI in education. He is also enthusiastic about strengthening industry–academic connections, frequently incorporating real-world examples and practices from industry to enrich teaching materials.

Puwadol Oak Dusadeerungsikul, Chulalongkorn University

Puwadol Oak Dusadeerungsikul, Ph.D., is an Assistant Professor in the Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Thailand. His research focuses on Operations Research, with emphasis on mathematical modeling, optimization, and algorithm development for complex decision-making problems. His work studies scheduling, routing, and resource allocation systems that arise in real-world applications such as logistics, agricultural, and healthcare systems. He received his Ph.D. in Industrial Engineering from Purdue University in 2020 and his M.S. in Industrial Engineering from the Georgia Institute of Technology in 2015.

Lauren Mullikin, Virginia Polytechnic Institute and State University

I am a junior in Industrial and Systems Engineering with a minor in Data and Decisions at Virginia Tech. I serve as an ISE Ambassador, am a member of Alpha Pi Mu, and currently serve as President of Institute of Industrial and Systems Engineers. My academic interests include data-driven decision-making, optimization, and systems improvement.

WIP: Strategies and Practices to Overcome Irresponsible and Low-Effort Deliverables (SPOILED) in an Excel-Based Database Class

abstract

While tools such as ChatGPT and Copilot have introduced new opportunities for teaching and learning in higher education, they have also become shortcuts for students seeking to bypass homework and take-home exams. A small but noticeable minority of students bypass assignments by copying questions into a chatbot and submitting the outputs with little or no learning or critical thinking. Such low-effort use of generative AI tools raises legitimate concerns about the rigor and validity of existing assessment designs. AI-generated answers often include details irrelevant to course content, non-functional formulas and code, incorrect variable names and references, or incorrect answers nearly matching ChatGPT's basic output. Such patterns are often distinctive enough for instructors to recognize, as the answers contain technical errors and inconsistencies rarely seen in authentic student work. To deter and detect such misuse, the development of the Strategies and Practices to Overcome Irresponsible and Low-Effort Deliverables (SPOILED, pun completely intended) framework was initiated. In its current stage, the framework assesses the vulnerability of Excel- and SQL-based assessments to low-effort misuse of AI tools. The framework also examines how much contextual data is needed for ChatGPT to produce acceptable answers. Initial findings indicate that Excel formula writing questions are vulnerable when students provide complete data tables, but can become resistant with small design choices such as shifting table addresses. SQL query writing questions are vulnerable when students provide ChatGPT with the database schema diagram. These findings support the use of course-specific formatting requirements as well as proctored quizzes and exams to verify student proficiency in Excel and SQL. Beyond these early results, the SPOILED framework provides instructors with a consistent approach to deter low-effort AI usage across different courses. Future work will extend the framework to question types used in a wider range of engineering education assessments. The study underscores the importance of systematic research on responsible AI use and the development of evidence-based methods to mitigate low-effort AI misuse in higher education.

keywords: generative AI in education, ChatGPT, responsible AI use, AI misuse, academic integrity, Excel-based assessment, SQL-based assessment

introduction

Generative artificial intelligence (AI) tools such as ChatGPT and Copilot have introduced new opportunities for both teaching and learning in higher education [1], [2], [3]. In technical courses, these tools can support student learning by providing rapid explanations, summarizing complex material, and generating example solutions that students can use to check their work. At the same time, the accessibility and convenience of generative AI tools can enable students to bypass the intended learning process, particularly in homework assignments and take-home exams [4]. Instructors have reported cases where students simply copy question text into ChatGPT and submit its output with little or no critical engagement [5]. This type of behavior raises significant concerns for academic integrity and can undermine the development of authentic problem-solving skills [6].

In this paper, *low-effort AI misuse (LAM)* is defined broadly to include both "zero-effort" and "low-effort" misuse of generative AI. *Zero-effort misuse* refers to copying and pasting AI-generated outputs without any learning, understanding, or modification. *Low-effort misuse* refers to submitting AI-generated work with minimal comprehension and without verifying that it follows the requirements of the course. In both cases, the student's work may appear complete, but the underlying learning goals, including careful interpretation of problem statements, error checking, and iterative refinement of formulas or code, are not achieved. The intent of this research is not to discourage learning-oriented AI use; rather, it focuses on identifying when AI outputs can be graded as fully or almost fully correct. Accordingly, this study evaluates only the chatbot's basic output rather than iterative, engaged prompting.

Importantly, LAM is used here to describe a specific assessment-use pattern and vulnerability condition, not a general characterization of students or their motivations. Many students use generative AI productively for brainstorming, explanation, feedback, verification, and skill development. Students may also use these tools responsibly to support professional development, academic research tasks, communication, and career preparation when they remain actively engaged in critiquing, revising, and verifying the output. This paper focuses only on cases where AI-generated outputs can be submitted with little verification while still meeting grading criteria. The central objective is therefore to help instructors identify which assessment items are vulnerable under realistic low-effort conditions and make targeted redesign choices that preserve learning goals.

The risks of LAM are especially relevant in courses involving Excel formulas and SQL query design, where many assessments require students to produce structured outputs with predictable formats. When sufficient context is provided, generative AI chatbots can often generate formulas and queries that are close to correct and appear course-like in form [7]. As a result, many traditional assessment items in these courses may be susceptible to low-effort AI misuse, particularly when students supply the chatbot with structured artifacts such as complete data tables, database schemas, or lecture notes. This creates two interconnected challenges for instructors: determining which assessment items are most vulnerable under realistic student behaviors, and redesigning those items to preserve learning outcomes and fair evaluation. However, it is often difficult to predict vulnerability in advance. Existing guidance frequently emphasizes broad policy recommendations or high-level redesign principles, but provides limited support for systematically testing specific assessment items against modern generative AI systems under controlled, low-effort conditions. This gap motivates the need for a practical framework that instructors can use to evaluate whether an assignment or exam item is vulnerable to LAM, and to support targeted redesign decisions when needed.

To address this need, this Work-in-Progress (WIP) paper presents the SPOILED framework (Strategies and Practices to Overcome Irresponsible and Low-Effort Deliverables). The SPOILED framework provides a systematic method to evaluate the vulnerability of Excel-based and SQL-based assessment items to low-effort AI misuse using ChatGPT. ChatGPT is selected in this study because it is the predominant generative AI tool used by undergraduate students, with one 2025 study reporting 93.8% of surveyed students using ChatGPT [8]. For each assessment item, ChatGPT is tested under increasing levels of contextual input, ranging from the question

text alone to the full set of relevant course materials. The resulting outputs are evaluated against instructor solutions using a rubric-based process to determine whether the item is vulnerable or resistant under low-effort conditions. The next section describes the course setting used in this study, the SPOILED framework, and the methodology used in the preliminary evaluation.

framework and methodology

study context and course setting

The course emphasizes hands-on data management and analytics using Excel and SQL and typically enrolls around 180 to 200 students per semester. Instruction is delivered primarily through Excel-based activities, with lecture sessions combining instructor-led explanations and live problem solving, during which students take notes and practice alongside prepared examples. Students complete homework assignments in Excel and submit final answers through Gradescope, which is used for submission and grading logistics because it supports efficient rubric-based feedback compared to the technical challenges of grading Excel files directly. In addition to regular homework, the course includes take-home exams structured as extended problem sets with less scaffolding than homework assignments.

Instructor observations and student comments in Spring 2025 indicated that a notable minority of students attempted to use ChatGPT, especially during high-workload periods of the semester. Based on a review of student submissions and grading patterns, the instructor estimated that at least ten percent of students, and possibly as many as one-third, relied on ChatGPT either in clear cases or in submissions raising strong suspicion; these figures should be interpreted as observational estimates rather than a direct measure of prevalence. Several submissions contained errors unlikely to occur in genuine student work, including formulas or queries that did not run, references to nonexistent columns or table names, incorrect primary key to foreign key relationships, and the use of functions or formula structures outside course material. The patterns were most noticeable during high-workload periods of the semester (for example, midterm weeks), consistent with the literature on generative AI usage during these periods [9]. Compared to earlier shortcutting approaches such as copying from peers or using solution sites (e.g., Chegg, Course Hero), modern chatbots can more readily adapt to changed question parameters and can generate course-like answers when sufficient context is provided [10]. Several suspicious submissions were also strikingly similar to outputs produced when only the basic question text was provided to ChatGPT, further motivating the need to evaluate assessment vulnerability under low-effort conditions.

SPOILED framework overview

The SPOILED framework (Strategies and Practices to Overcome Irresponsible and Low-Effort Deliverables) is designed to systematically evaluate the vulnerability of Excel-based and SQL-based assessment items to low-effort AI misuse (LAM). The primary goal is to determine how much contextual information needs to be provided before ChatGPT can produce answers that are correct and formatted according to course requirements. By identifying these thresholds, the framework provides a structured way to examine whether a question is vulnerable to LAM. An overview of the SPOILED framework is provided in Figure 1.

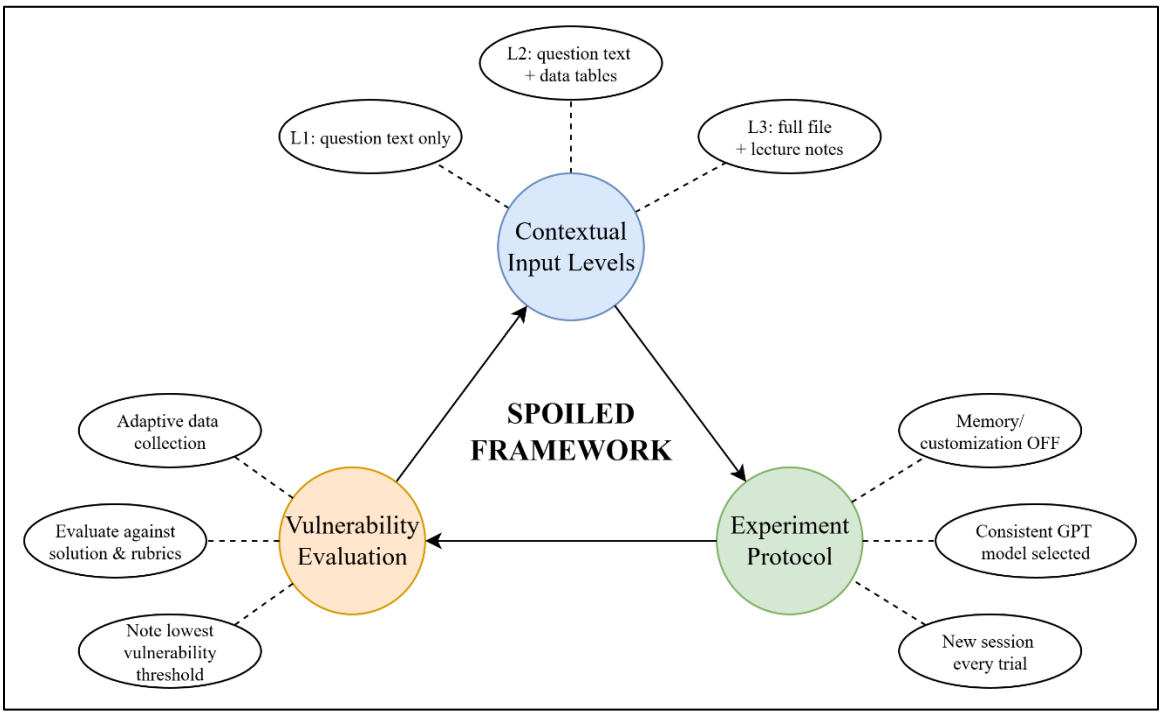

Figure 1. Overview of the SPOILED framework

The framework consists of three components, which are described in the following subsections: contextual input levels, a controlled experimental protocol, and rubric-based vulnerability evaluation. Because LAM is defined as shortcutting with minimal understanding, the framework evaluates only ChatGPT's initial output rather than iterative, learning-oriented prompting.

contextual input levels

A key component of the SPOILED framework is the systematic testing of assignment questions against ChatGPT using different amounts of contextual information. This design reflects how students may realistically attempt to use generative AI tools, ranging from minimal prompts using only question text to providing the chatbot with nearly complete course materials. For this study, three distinct levels of contextual input were defined.

Level 1 (Question text only, missing key context and data): Only the basic question text is copied and pasted into the chatbot. This simulates students seeking quick answers with no added context, typically resulting in incomplete, incorrect, or generic responses.

Level 2 (Question text plus surrounding context and data): The question text is provided along with the surrounding context, instructions, and data from the assignment. This simulates students giving just enough detail to improve ChatGPT's response, but still without the lecture notes that may contain key requirements not covered by the expanded question text, context, and data.

Level 3 (Full assignment file and lecture notes): The complete homework or take-home exam worksheet is uploaded along with the relevant lecture notes. In this course, homework

worksheets are placed side by side with lecture notes, while take-home exams are provided in separate Excel files. At this level, ChatGPT has nearly all of the technical inputs needed to construct a correct answer, which can increase accuracy. However, the larger volume and complexity of context can also overwhelm the model, leading it to "overthink" and propagate errors through its reasoning chain [11].

The three levels broadly capture the range of low-effort misuse scenarios relevant to this study and form the basis for the experimental protocol described in the next section.

experimental protocol

To ensure consistency in testing, a new, blank ChatGPT session is started for each trial. At the time of the experiment (January 2026), ChatGPT's response to a prompt is informed by the prior messages within that same session, so beginning each trial in a fresh session helps prevent carryover from earlier prompts, intermediate outputs, or previously tested items. All tests in this study are conducted using ChatGPT 5.2 in Auto mode, the system's default setting available at the time of the experiment. This choice is intentional, as the preliminary stage of the research aims to establish baseline results, and the default mode aligns most closely with the definition and scope of low-effort AI misuse (LAM).

It is noted that ChatGPT provides additional customization options, particularly through the Personalization settings and the Custom Instructions feature; however, these should be disabled for all sessions to maintain consistency and reduce variation across trials. Specifically, Custom Instructions are turned off, and the settings to reference saved memories and to reference chat history are disabled, so the system relies only on the information intentionally provided within the current trial. Before each trial, the chatbot is reloaded and the Personalization settings are verified to be blank, as shown in Figure 2.

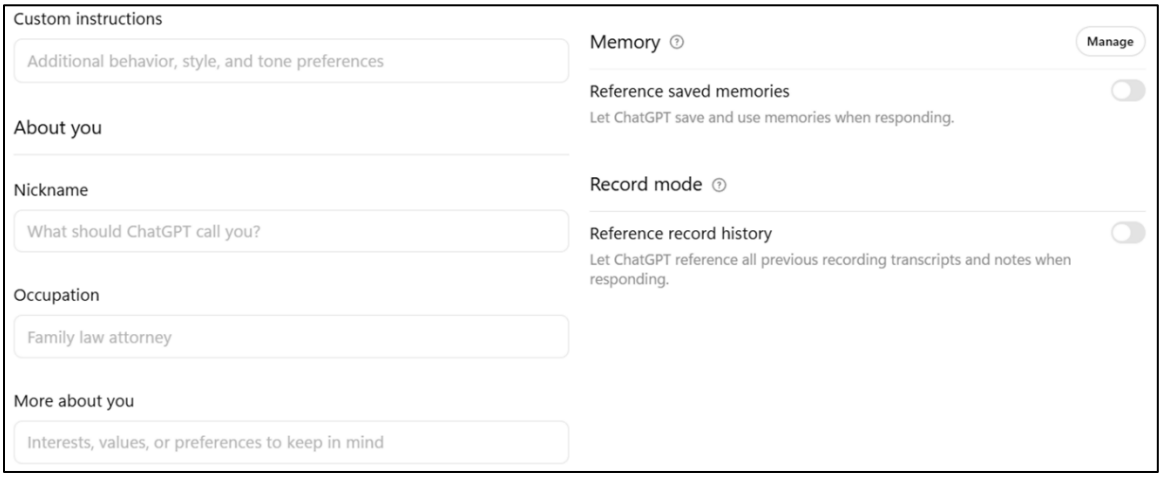

Figure 2. ChatGPT personalization and memory settings disabled for experimental trials

All tests in this study are conducted using the desktop version of ChatGPT. Because the underlying model is the same across platforms (Desktop vs browser vs mobile) and provides a final answer regardless of interface, the outputs are essentially similar, and the choice of platform

should not affect the results being considered. To ensure consistency across repeated trials and to reduce the likelihood that usage limits would interrupt testing, the experiments are conducted using a ChatGPT Plus subscription account (listed at $20 per month at the time of the experiment). Many college students subscribe to ChatGPT Plus for homework assignments either by purchasing a monthly subscription themselves or by sharing accounts with peers [2]. A paid subscription can provide more consistent access and reduce plan-related constraints that could otherwise introduce noise into repeated, multi-trial experimental runs, and future work can expand on the comparison between the free and paid versions of ChatGPT.

vulnerability evaluation

Each assessment item is tested across three levels of contextual input, as described in the previous subsection. For every trial, the chatbot's outputs are recorded and then compared against the instructor's correct solution and grading rubrics. The rubrics emphasize correctness, completeness, and adherence to course-specific requirements. This systematic comparison allows the framework to capture whether additional context improved, degraded, or otherwise altered the quality of ChatGPT's responses. In this study, each combination of assessment item and contextual input level is tested at least three times. If at least two out of three ChatGPT outputs scored sufficiently high based on the rubrics given a certain level of contextual input (for example, Level 1), the assessment is evaluated as vulnerable at that level of contextual input (L1-vulnerable).

For each level (L1-L3), an assessment item is classified as *L1/L2/L3-vulnerable* if ChatGPT's basic output at that level meets the rubric criteria for correctness, completeness, and adherence to course-specific requirements and formatting; otherwise, it is classified as *L1/L2/L3-resistant*. An assessment item is *LAM-resistant* if it is resistant at all three levels. During preliminary testing, non-monotonic patterns were observed; for example, Q3 defined below was L2-vulnerable while L3-resistant.

sample question set

To illustrate the application of the SPOILED framework, a sample set of assessment items was selected from the course. These questions represent typical tasks in Excel-based and SQL query writing assessments.

Q1. Excel VLOOKUP formula writing: Using the Normal Ranges of a provided Excel table, write a VLOOKUP formula according to the class's formula structure to return Customer Name for Project ID #4. The table referenced in this question is 26 columns by 364 rows (including the header row), with the top-left cell at B2.

> *Level 1*: Only the immediate question text is provided. The Excel table is then copied and pasted into ChatGPT in the same message. By default, pasting an Excel table into ChatGPT provides both a tab-delimited text version of the data and an image snapshot.
> *Level 2*: Same as above, except that the immediate question text and additional instructions are both provided at the start.
> *Level 3*: The question text and additional instructions are provided. The Excel file containing both the lecture notes and the homework file is uploaded into ChatGPT.

Q2. SQL 5-Table INNER JOIN query writing: Identify the correct five-table join order and write the INNER JOIN query to join the tables, given the class's MS Access database.

Level 1: Only the immediate question text is provided.
Level 2: The immediate question, additional instructions, and the SQL database schema are uploaded.
Level 3: The question text and additional instructions are provided. The Excel file containing the lecture notes, the SQL database schema diagram, and the homework file is uploaded into ChatGPT.

Q3. Excel sheet navigation multiple-choice: Using the provided Grade Calculator Excel file, navigate, modify the input values as instructed, and compute the required outputs.

Level 1: Only the immediate question text is provided.
Level 2: The immediate question, additional instructions, and the cells in the Grade Calculator Excel file are copied and pasted directly into ChatGPT (but the file is not uploaded).
Level 3: The immediate question, additional instructions, and the Grade Calculator Excel file are uploaded.

Together, these questions encompass a substantial segment of the Excel and SQL assessments used in the course selected for this study and provide a representative basis for evaluating vulnerabilities within the SPOILED framework.

preliminary results and discussion

The SPOILED framework was applied to the three questions (Q1, Q2, and Q3) described above. Each question was tested under three levels of contextual input, and each question-level combination was evaluated in at least three independent trials. Results are summarized with accompanying observations and are discussed in the following section.

<u>results of Q1. Excel VLOOKUP formula writing</u>

Across trials, the Level 1 and Level 2 answers are largely similar:

=VLOOKUP(4,OuterJoinView1!$E:$J,6,FALSE)

The differences become more evident at Level 3, where the returned formulas are more variable and include multiple inconsistent range specifications, for example:

=VLOOKUP(4, OuterJoinView1!F2:K100, 6, FALSE)
=VLOOKUP(4, OuterJoinView1!A:B, 2, FALSE)
=VLOOKUP(4, OuterJoinView1!F:K, 6, FALSE)

The correct answer for this question is either:

=VLOOKUP(4, OuterJoinView1!$F$2:$K$364, 6, FALSE)
=VLOOKUP(4, OuterJoinView1!$F$2:$K$364, 6, 0)

The second argument of the formula must use absolute references (F4) for the cell references, and must mention the row addresses.

Under the rubric used in this study, all ChatGPT-provided outputs for Q1 are incorrect relative to the instructor solution and do not adhere to the course-specific formula structure requirements. Following the SPOILED framework definitions, Q1 is therefore classified as L1-resistant, L2-resistant, and L3-resistant, and is overall LAM-resistant. It is also noted that, although the Level 1 and Level 2 formulas do not follow course-specific requirements (for example, the second argument referencing entire columns rather than the required Normal Ranges), these outputs may appear plausible and may not be readily distinguishable from genuine student work. At Level 3, the model appears to become overloaded by the volume and complexity of the uploaded file, leading to greater inconsistency in the selected lookup ranges and column indices across trials.

Finally, sensitivity to minor worksheet structure was observed: if the referenced table's top-left cell were located at A1 rather than B2, the Level 1 and Level 2 outputs would have been correct in structure, suggesting that ChatGPT implicitly assumes that the OuterJoinView1 table begins at A1. In that counterfactual case, Q1 would have been classified as L1-vulnerable and L2-vulnerable, indicating that a small modification to worksheet layout can materially change an assessment item's vulnerability to LAM.

results of Q2. SQL 5-Table INNER JOIN query writing

Level 1's output (same result across all three trials):

```
SELECT *
FROM Projects
JOIN Customers ON Projects.customer_id = Customers.customer_id
JOIN Operators ON Customers.operator_id = Operators.operator_id
JOIN Parts ON Operators.part_id = Parts.part_id
JOIN Machines ON Parts.machine_id = Machines.machine_id;
```

Under the rubric used in this study, the Level 1 output does not adhere to the course's MS Access query conventions and also exhibits schema-related issues, including inconsistent column naming and incorrect or unsupported relationship assumptions. As a result, Q2 is classified as L1-resistant.

At Level 2, the outputs are mixed. Two trials returned the same general structure, while the third trial produced a variant that listed all column names verbatim rather than using SELECT *, and in some cases introduced additional joins or relationships beyond what was required. An example Level 2 output is:

```
SELECT *
FROM (((Projects
INNER JOIN Customers ON Projects.CustomerID = Customers.CustomerID)
INNER JOIN Operators ON Projects.OperatorID = Operators.OperatorID)
INNER JOIN Parts ON Projects.ProjectID = Parts.ProjectID)
INNER JOIN Machines ON Operators.MachineID = Machines.MachineID;
```

The Level 2 outputs more frequently resemble MS Access SQL structure (for example, use of INNER JOIN and nested parentheses). As a result, Q2 is classified as L2-vulnerable overall. It is

noted that the reliability of ChatGPT output is significantly improved once the database schema is provided (Figure 3).

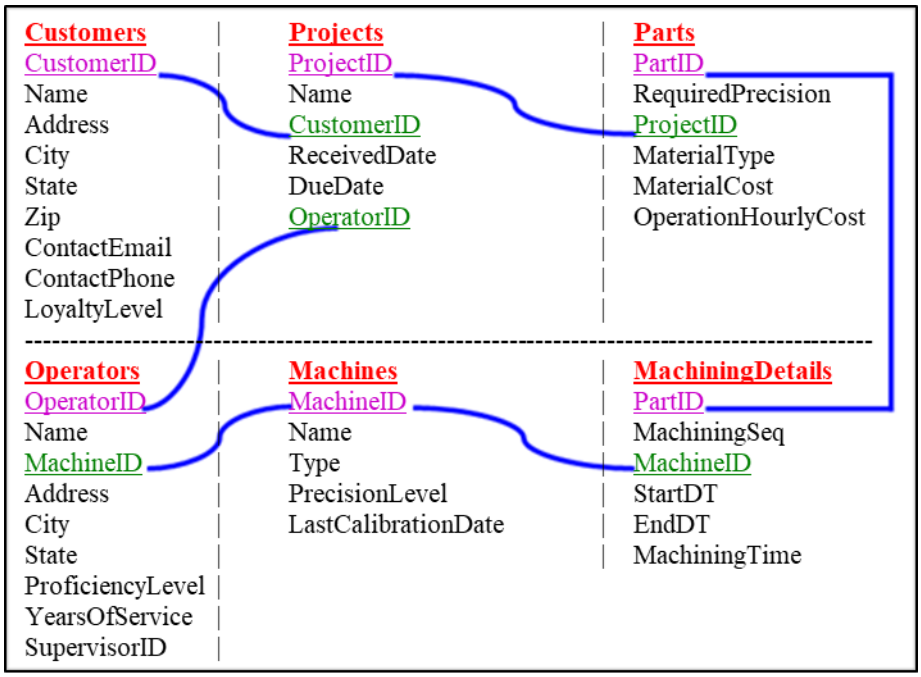

Figure 3. Database schema diagram for Q2

At Level 3, the outputs are mostly incorrect, with one example:

```
SELECT *
FROM ((((Projects
INNER JOIN Customers ON Projects.CustomerID = Customers.CustomerID)
INNER JOIN Operators ON Customers.OperatorID = Operators.OperatorID)
INNER JOIN Parts ON Operators.PartID = Parts.PartID)
INNER JOIN Machines ON Parts.MachineID = Machines.MachineID);
```

The Level 3 outputs are incorrect under the rubric due to schema and naming inconsistencies (for example, mismatching primary key and foreign key relationships). Furthermore, the output queries do not run in the course's MS Access database. Following the SPOILED framework definitions, Q2 is therefore classified as L3-resistant.

results of Q3. Excel sheet navigation multiple-choice

At Level 1, ChatGPT does not provide answer-choice selections for this item. Since this version of ChatGPT does not generate multiple-choice answer options by default under the Level 1 input condition, Level 1 is treated as not applicable for Q3, and interpretation focuses on Levels 2 and 3.

Across trials, the Level 2 outputs are consistent, and ChatGPT selected the correct choices in all three trials:

( ) 87.272727 and ( ) 375

The differences become more evident at Level 3. Across three trials, the selected choices vary and include both correct and incorrect selections.

One incorrect and one correct choice: ( ) 88.872727 and ( ) 375
Both choices are incorrect: ( ) 89.672727 and ( ) 405
One incorrect and one correct choice: ( ) 87.272727 and ( ) 405

Under the rubric used in this study, Q3 is therefore classified as L2-vulnerable, because the Level 2 outputs consistently meet the rubric criteria by selecting the correct choices under low-effort conditions. In contrast, Q3 is classified as L3-resistant, because the Level 3 outputs are not consistently correct across trials and frequently include incorrect selections. This pattern also provides an example of a non-monotonic relationship across contextual input levels, in which additional uploaded context does not necessarily improve performance and may instead reduce response reliability for this item.

summary of preliminary results

Table 1 summarizes the vulnerability classifications for the three preliminary assessment items and highlights that vulnerability should be evaluated separately at each contextual input level.

Table 1. Summary of Preliminary Results

Question	Description	L1	L2	L3
Q1	Excel VLOOKUP formula writing with table starting at cell B2	*LAM-resistant*		
Q1*	Excel VLOOKUP formula writing with table starting at cell A1	*L1-vulnerable*	*L2-vulnerable*	*L3-resistant*
Q2	MS Access SQL join query with the required join order	*L1-resistant*	*L2-vulnerable*	*L3-resistant*
Q3	Multiple-choice numerical selection item from navigating an Excel file	N/A (choices not generated)	*L2-vulnerable*	*L3-resistant*

implications for assessment design and the SPOILED framework

The preliminary results highlight two overarching patterns. First, the vulnerability of an assessment item is influenced by the level of contextual inputs provided to ChatGPT. Second, the relationship between contextual input level and vulnerability is not necessarily monotonic. Assessment items can be L2-vulnerable yet L3-resistant, indicating that additional context may not reliably improve the chatbot's basic output.

For Excel formula-based items, the Q1 versus Q1* comparison suggests that small, low-cost worksheet design choices can materially improve LAM-resistance without requiring major changes to learning objectives. Minor layout changes such as shifting a table's top-left cell away from A1 (for example, starting at B2) and selectively hiding rows or columns can reduce the likelihood that ChatGPT will infer correct table boundaries using default assumptions. These are

actionable redesign options that can be implemented with minimal disruption to the assessment prompt and grading rubric.

In addition, course-specific requirements appear to provide a meaningful layer of resistance against LAM. Requirements such as using Normal Ranges, fully locking cell references, and explicitly referencing row addresses require specificity that the model often omits or generalizes (for example, referencing entire columns). As a result, even when the chatbot produces a formula structure that may return the correct value, the output can still fail to meet course expectations. This reinforces that LAM-resistance is not only a matter of whether an answer is "correct," but whether it satisfies course requirements.

For SQL query-based items, the diagram of the database schema is observed to be a major driver of LAM-vulnerability. When the diagram is not provided and the question requires more than basic inference, ChatGPT frequently returns incorrect results. When the diagram is available, ChatGPT can often infer table relationships from relationship lines and column naming cues, and thus can generate executable MS Access queries that are frequently close to correct. Within this study, several obfuscation tactics were attempted, including adding background noise to the schema diagram (for example, additional lines), but the current version of ChatGPT overcame these attempts with little difficulty. While not providing the database schema diagram can improve resistance against LAM, the diagram directly supports learning, particularly for students who are still developing relational reasoning and join logic. This creates an assessment design tension: not providing the database diagram may reduce LAM, but it may also increase unnecessary friction for learning. One potential relief is to shift the most LAM-vulnerable components of the task into timed, in-class assessments where students must interpret the schema and produce working formulas or queries, while reserving take-home work for practice, preparation, and feedback.

Across all assignment types, increasing the amount of contextual information available to the chatbot through file uploading does not necessarily increase LAM-vulnerability. Although Level 3 provides substantially more information than Levels 1 and 2, LAM is defined as shortcutting with minimal understanding and minimal follow-up. Under these conditions, larger and more complex context can require the user to engage more actively with the chatbot to locate relevant information, resolve ambiguities, and correct inconsistencies. As a result, additional context may not translate into increased LAM-vulnerability and may instead reduce output reliability for some items, as observed in Q3. These findings also support key design choices in the SPOILED framework. In particular, they justify treating vulnerability levels as independently classified rather than assuming a monotonic increase with additional context.

It is also noted that incorrect ChatGPT-like outputs do not necessarily prove the presence of LAM in student submissions, since genuine student errors can occur and may resemble chatbot errors. However, incorrectness can still serve as a practical deterrent to LAM because it reduces the likelihood of receiving full credit while increasing the perceived risk of academic integrity investigation. In this sense, LAM-resistance can support both assessment validity and academic integrity, even when it cannot be used to conclude AI misuse with certainty.

Interpretation-oriented and reflective questions were also explored in this study, consistent with common instructional guidance for reducing AI-enabled shortcutting. These included questions asking students to explain query logic and interpret outputs, but these questions were not reported here for brevity. Following the SPOILED framework, these questions are mostly L1-vulnerable and L2-vulnerable, as ChatGPT was frequently able to generate coherent explanations and interpretations, suggesting that interpretive framing alone does not reliably reduce LAM-vulnerability. In addition, these open-ended items introduce significant grading logistics challenges in a large-enrollment setting. Compared to strictly specified formulas or queries, interpretation responses are more difficult to scaffold with clear instructions and typically required more complex rubrics to grade consistently, increasing grading time and inconsistency. These observations further motivate the use of a structured, item-level vulnerability evaluation approach in the SPOILED framework.

conclusion and future work

This study is preliminary and exploratory. Only a limited set of questions is presented here for brevity, and the findings should be interpreted as early evidence rather than general conclusions. Future work will expand the Excel and SQL question bank to include items requiring debugging, validation, interpretation, and justification beyond direct formula or query generation. Additional work may also compare input modalities, model versions, platforms, and access tiers, including free and paid versions of ChatGPT. Finally, future studies should examine redesign strategies that balance educational value with improved LAM-resistance.

This Work-in-Progress paper introduced the SPOILED framework as a systematic method for evaluating low-effort AI misuse vulnerability in Excel-based and SQL-based assessment items across multiple contextual input levels. The preliminary results demonstrate that vulnerability can be strongly driven by the contextual input provided to the chatbot, particularly database schemas for SQL, and that vulnerability does not necessarily increase with additional context. Small, low-cost design choices, including worksheet layout and explicit formatting requirements, can materially change whether an item is vulnerable or resistant under low-effort misuse.

Even as generative AI tools become increasingly common, students still need the foundational ability to read problem statements carefully, verify outputs, and write and revise formulas or code on their own [12]. These foundations build the technical judgment required for advanced coursework and professional practice [13], particularly in engineering contexts where problems are less structured, involve messier constraints, and require diagnosis and validation rather than simply producing a final answer [14]. The SPOILED framework is intended to support responsible AI use in technical coursework by helping instructors identify vulnerable items systematically and redesign assessments to remain valid and fair as AI tools improve. LAM is therefore not only an academic integrity issue, but also a barrier to preparing students for future coursework and professional expectations in industry.

references

[1] M. I. Baig and E. Yadegaridehkordi, “ChatGPT in the higher education: A systematic literature review and research challenges,” *Int. J. Educ. Res.*, vol. 127, p. 102411, Jan. 2024, doi: 10.1016/j.ijer.2024.102411.

[2] F. Wu, Y. Dang, and M. Li, “A Systematic Review of Responses, Attitudes, and Utilization Behaviors on Generative AI for Teaching and Learning in Higher Education,” *Behav. Sci.*, vol. 15, no. 4, p. 467, 2025.

[3] K. Hon, “Generative AI in higher education: A systematic review of its effects on learning outcomes and academic performance,” *J. Educ. Technol. Syst.*, p. 00472395251400089, 2025.

[4] D. R. Cotton, P. A. Cotton, and J. R. Shipway, “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,” *Innov. Educ. Teach. Int.*, vol. 61, no. 2, pp. 228–239, 2024.

[5] W. Kerney, “Treachery and Deceit: Detecting and Dissuading AI Cheating,” *J. Comput. Sci. Coll.*, vol. 40, no. 9, pp. 10–17, 2025.

[6] A. M. Hasanein and A. E. E. Sobaih, “Drivers and consequences of ChatGPT use in higher education: Key stakeholder perspectives,” *Eur. J. Investig. Health Psychol. Educ.*, vol. 13, no. 11, pp. 2599–2614, 2023.

[7] S. R. Piccolo, P. Denny, A. Luxton-Reilly, S. H. Payne, and P. G. Ridge, “Evaluating a large language model’s ability to solve programming exercises from an introductory bioinformatics course,” *PLOS Comput. Biol.*, vol. 19, no. 9, p. e1011511, 2023.

[8] V. Morell-Mengual, O. Fernández-García, C. Berenguer, J. Ortega-Barón, M. D. Gil-Llario, and V. Estruch-García, “Characteristics, motivations and attitudes of students using ChatGPT and other language model-based chatbots in higher education,” *Educ. Inf. Technol.*, pp. 1–18, 2025.

[9] M. Abbas, F. A. Jam, and T. I. Khan, “Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students,” *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, p. 10, 2024.

[10] T. Susnjak and T. R. McIntosh, “ChatGPT: The end of online exam integrity?,” *Educ. Sci.*, vol. 14, no. 6, p. 656, 2024.

[11] M. A. Ferrag, N. Tihanyi, and M. Debbah, “Reasoning Beyond Limits: Advances and Open Problems for LLMs,” Mar. 26, 2025, *arXiv*: arXiv:2503.22732. doi: 10.48550/arXiv.2503.22732.

[12] Y. Walter, “Embracing the future of Artificial Intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education,” *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, p. 15, 2024.

[13] V. Edmondson and F. Sherratt, “Engineering judgement in undergraduate structural design education: enhancing learning with failure case studies,” *Eur. J. Eng. Educ.*, vol. 47, no. 4, pp. 577–590, 2022.

[14] M. Pereira Pessoa, “Guidelines for teaching with ill-structured real-world engineering problems: insights from a redesigned engineering project management course,” *Eur. J. Eng. Educ.*, vol. 48, no. 4, pp. 761–778, 2023.