UDC 378.091.26:004.8

Borys Kuzikov
Sumy State University, Sumy, Ukraine
ORCID ID 0000-0002-9511-5665
b.kuzikov@dl.sumdu.edu.ua

Svitlana Kuzikova
Sumy State Pedagogical University named after A. S. Makarenko, Sumy, Ukraine
ORCID ID 0000-0003-2574-9985
kuzikova_svetlana@ukr.net

THE IMPACT OF LARGE LANGUAGE MODELS ON THE VALIDITY OF THE WRITTEN LABORATORY REPORT AS AN ASSESSMENT TOOL FOR LEARNING OUTCOMES IN HIGHER EDUCATION

Abstract. Written laboratory reports remain a principal assessment instrument in distance education, yet the growing capability of large language models (LLMs) undermines the tacit assumption on which the validity of such reports rests: that writing the report requires first solving the task. This article examines whether leading current-generation LLMs, given a minimal prompt, can fully complete typical laboratory assignments of a university Databases course at the level of the highest student grade under the course's unmodified rubric. In a pilot case study with comparative benchmarking, eleven models from five providers (Anthropic, OpenAI, Google, xAI, DeepSeek) were tested in their native chat interfaces on authentic assignments from a live course: closed tasks with deterministic outcomes and an open-ended experiment-design laboratory. All anonymised responses were rated manually by three independent raters. The current generation handles the content confidently: median correctness is 91.6%, four models reach ≥ 98.9%, and over half of all content scores fall within the top rubric band. However, solving the task and producing a submission-ready report proved to be distinct capabilities: every model's formatted MS Word report was rated below its content correctness, and two models returned no usable file at all. The remaining barrier is technical rather than cognitive: an agentic tool with access to the course files generated complete reports with genuine execution output for all 11 assignments of a semester in a single run. We conclude that, for the content component, the unsupervised written report no longer measures the student's skill. The defensive barriers of the written report examined in the study are ordered by their observed resilience: from those already fallen (the creativity of the open-ended task, assignments on the student's "own data") through those holding only against an isolated chat (linked assignments, the complex report template) to a potentially durable class of artefacts external to the model, which rests not on the models' cognitive limits but on their lack of authority over external state, and therefore does not expire with the next model generation.

Keywords: assessment validity; large language models; academic integrity

1. INTRODUCTION

Problem statement. For decades, the written laboratory report has served a dual function in technical education: it compelled the student to work through the full cycle of solving a task, and it gave the instructor grounds for judging whether the skill had been acquired. Both functions rested on a tacit assumption: to write the report, one must first solve the task. Current-generation large language models (LLMs) call this assumption into question. Their effect goes beyond helping the student find a reference or fix a syntax error faster: they are potentially capable of completing every step from the task statement to the formatted document without the student’s participation. If this capability is confirmed, the written report ceases to measure what it was created to measure. We term this phenomenon the validity erosion of assessment under the influence of artificial intelligence.

© Borys Kuzikov, Svitlana Kuzikova, 2026.

This vulnerability is systemic: tests and written reports dominate distance education as the cheapest formats to scale, so the loss of their validity affects the entire organisation of distance assessment. For the first of these instruments the process has already been documented: a scoping review [1] covering 53 studies and almost fifty thousand multiple-choice questions showed that free models of the previous generation failed most higher-education examinations, whereas GPT-4 passed most of them on a par with human examinees — leading its authors to conclude that summative testing can no longer be conducted without restricting access to such tools. This article examines whether the same fate has befallen the second of these instruments, the written laboratory report. An affirmative answer would mean that the report can no longer distinguish an acquired skill from generated text, irrespective of how many students actually exploit this possibility; whether skill acquisition itself has declined would require a longitudinal design and remains outside the scope of this work.

The simplest response, banning LLMs, fails for two reasons. The first is practical and officially acknowledged: the joint recommendations of Ukraine's Ministry of Education and Science and Ministry of Digital Transformation describe a ban on generative AI as "tantamount to the bans once proposed on calculators or computers" [2]: fluency with generative tools has itself become a labour-market skill. The second reason runs deeper, and the calculator analogy conceals it: a calculator performs a mechanical sub-operation within a task the student holds in mind as a whole, whereas an LLM past a certain capability threshold can take over the task in its entirety [3]. In the terms of cognitive offloading [4], this is the boundary beyond which a tool ceases to support thinking and begins to substitute for it, performing the work in the student's stead, outside the skill-acquisition loop. Educational practice is already attempting to hold this boundary, without evidence that the line it draws can be held: four of the five laboratory assignments used in this study carry a blanket prohibition ("The use of generative networks for completing the assignments is prohibited"), while the fifth draws the line more finely: "You may use AI as a search engine or as a means of preparing data. However, the design and conduct of the experiment and the analysis of its results must be carried out exclusively by you in person." Both policies assume that the boundary between "AI as a reference tool" and "AI as an executor" can be drawn and held. This study puts that assumption to an empirical test.

Analysis of recent research and publications. The theoretical vocabulary for such a test is provided by [5]: cheating should be treated not as a moral transgression but as a problem of assessment validity — tasks that "rely on students not using artificial intelligence, yet are unable to prevent them from doing so" are of little use for consequential assessment, regardless of whether the workaround qualifies as misconduct. A related conclusion is reached in [6]: if AI completes a task in a way that misleads an expert assessor, the problem lies not in the student's cunning but in the intellectual triviality of the task itself. This line has become institutionalised: the Artificial Intelligence Assessment Scale (AIAS) [7] deems the "no AI" level fit only for a controlled, in-person environment, and in its second edition the authors state that declaring an assessment "AI-free" without environmental control has become untenable, and reliable verification of authorship unachievable [8]. This is consistent with the low and unstable accuracy of software AI-text detectors, which simple paraphrasing defeats [9]. The Ukrainian ministerial recommendations [2] reproduce this scale verbatim and advise compensatory measures: oral interviews, projects in place of written assignments, tasks based on personally collected data. In the chain "validity theory → international framework → national policy" one link is missing: an empirical measure of how far the written report is already unverifiable. All these documents assume it; none measures it. This study supplies that measure on authentic laboratory assignments from a live technical course.

Empirical precedents trace a common trajectory. Study [10], using blind assessment by three independent raters (ICC = 0.932) on programming-course assignments, recorded the cross-generation dynamics: GPT-3.5 scored 30.9–48.8% depending on prompt quality, GPT-4

71.9–81.1%, and students averaged 91.9%; the gap narrowed but, as of GPT-4, had not closed. Study [11] tested GPT-3.5 on assignments from ten disciplines across seven Australian universities (including "code submission" and "experiment-based written assignment", structurally identical to our material) and explicitly called for the benchmarking to be repeated on future models. The authors of [12] concluded that even for GPT-3.5/GPT-4-era models it "may be sufficient for students simply to enter the assignment text and the rubric into these tools to obtain satisfactory or high marks". Our design brings together conditions never yet combined in these precedents: a current (2025–2026) model generation, blind assessment by several independent qualified raters with inter-rater reliability statistics, and a task range spanning closed problems with a single correct answer to open-ended experiment design.

In the Ukrainian segment of the literature, the problem has so far been approached mainly through perception and normative regulation rather than the models' empirical capability: violations involving LLM-generated texts still lack established terminology [13]; 45.9% of surveyed students cannot say with certainty whether LLM use constitutes an integrity violation [14]; students perceive ChatGPT as a reference-and-information tool [15], while the actual capability of the models may have advanced far beyond that role. The empirical line at the intersection of LLMs and assessment is only beginning to form: multi-model evaluation of chatbots in deep-search mode [16], comparison of AI-generated and instructor-designed teaching materials [17], psychometric validation of LLM-generated test items [18]. The capability of models to complete technical laboratory assignments in full has not yet been investigated within this line; this is the subject of the present article.

Aim and objectives of the study. The aim is to establish, on authentic assignments from a live course, whether leading current LLMs, given a minimal-effort prompt, can fully complete typical laboratory assignments in the Databases discipline. Three objectives serve this aim: (1) to determine the share of assignments that each tested model completes fully and without corrections; (2) to establish whether this capability is a consistent property of the current LLM generation as a whole or specific to individual models and providers; (3) to ascertain whether work completed by an LLM falls, under the course's current unmodified rubric, within the range of the highest student grade. That LLMs can complete some university assignments was established by our predecessors [10–12]; our contribution is a test of whether the gap that in 2023–2024 was still narrowing [10] has now been closed.

2. THE THEORETICAL BACKGROUNDS

The first theoretical anchor is the thesis of [5] that validity is a property not of an instrument but of a particular act of assessment: a task loses validity at the moment there appears an accessible way to complete it without demonstrating the expected competence. Task design shapes student behaviour: a closed problem with a single verifiable answer creates different incentives from an open-ended task requiring interpretation and justification, a distinction mirrored in our division of the material into closed tasks with deterministic outcomes and an open-ended experiment-design problem.

The second anchor is the revised taxonomy of educational objectives [19]. Study [12] used it to record the characteristic profile of 2023-era models: confident performance at the lower and middle levels, failure at the "create" level. Our design reproduces this logic on database material: executing a query with a deterministic result (Group A) corresponds to the "remember–apply" levels; independently designing and conducting an experiment with comparison of alternatives (Group B) to the "analyse–evaluate" levels; synthesising a ready-to-submit artefact to the "create" level. This mapping establishes at which cognitive level, if any, the model still lags behind the student. If the current generation covers the upper levels too, the recommendation to "move assessment up to the create level" [12] becomes obsolete

less than three years after it was made. The taxonomy-extension study [20] adds the observation decisive for our argument: all the productive practices of student–model interaction it documents (critical selection, iterative refinement, prompt formulation) are invisible in the final written artefact. That study captures the upper bound of LLM use (deliberate, reflective engagement); our experiment measures the lower bound (a minimal prompt with no refinement). If both bounds yield an indistinguishable artefact, the report cannot discriminate between the best and the worst scenarios of use, which is the operational definition of the instrument's validity erosion.

The third strand concerns the reliability of model outputs, which is itself a validity problem: a generated output looks the same to the assessor whether or not it is factually correct. In Group B, this is a concrete risk: a generated analysis of a query execution plan is equally convincing whether the model genuinely reproduces the logic of EXPLAIN output or assembles a plausible sequence of assertions, and this analysis is what the rubric treats as the marker of the highest level of student competence. We use the established term hallucination for factually false outputs, acknowledging the argument of [21] that it anthropomorphises a process indifferent to the truth of its output. The reliability picture is changing rapidly: generated-reference accuracy in chat mode in 2024–2025 ranged from 15% to 59% [22], whereas current-generation chatbots in deep-search mode reached 100% with no fabricated references [16]; the leap comes not only from new generations but from modes with tool access; our design accounts for this with a separate agentic arm.

3. RESEARCH METHODS

The study is designed as a pilot case study with comparative benchmarking: the responses of eleven leading current LLMs to authentic laboratory assignments from a live course are assessed manually by three independent raters against the same rubric applied to student reports. The material comprises five laboratory assignments from the course "Databases and Information Systems" as taught at Sumy State University in 2024–2026. The case has a substantive rationale: owing to the security situation, this border-region university operates fully remotely (including for roughly 20% of its students currently abroad), so the in-person proctored or oral assessment that the AIAS framework [7, 8] requires for its "no AI" level is in practice unavailable, and the validity of remote written formats acquires heightened importance. The course spans two semesters: the first ("SQL Fundamentals", 11 laboratory assignments) builds the basic skill of writing queries; the second builds a deeper understanding of how databases operate (performance, indexes, integrity). The AI policy is tied to task type, not semester: the assignments with deterministic outcomes (Group A) carry the blanket prohibition quoted in the Introduction, the open-ended performance laboratory (Group B) carries the finer-grained one.

Group A comprises four closed tasks with a deterministic expected result: subqueries, table joins, grouping and aggregation, and working with partially unique data; the first three use the classic SCOTT teaching schema, the fourth data generated by the student or real anonymised data. The SCOTT schema has circulated in open sources for decades, so a model may reproduce the simplest queries from its memorised corpus; we deliberately kept these assignments because the object of study is not SQL syntax but the completeness of the assignment as an integral submittable artefact, including atypical subtasks for which direct recall is unlikely.

Group B is a laboratory assignment containing nine real performance problems involving large data volumes (hundreds of thousands to a million rows), which in its full set unfolds into 45 variants (nine problems × five DBMSs); the study covers three, each on a different DBMS (PostgreSQL, IBM Db2, MS SQL Server), preserving the diversity of problem types and

execution environments without combinatorial expansion. This assignment is the direct test of the boundary that its own policy attempts to draw. The rubric assesses less the fact of solving the problem than the completeness with which the process is documented: data preparation, conduct and analysis of the experiment, alternatives considered. The highest mark requires justifying why the chosen solution is superior to the alternatives – a demand answerable only by comparative judgement, the ability that separates "apply" from "analyse" and "evaluate" [19]. The assessors apply this rubric to the LLM responses without modification.

Operationally, the Group B rubric is decomposed into five subscales – one per stage of the work and type of submittable artefact: (1) input data; (2) conduct of the experiment; (3) report (Microsoft Word, .docx); (4) presentation script; (5) slides (Microsoft Power Point, .pptx). Each subscale is rated ordinally, 0–5: 0 – the artefact is absent; 1–2 – requires substantial rework; 3–4 – could be submitted with minor corrections for a partial mark; 5 – on a par with the best students, no rework needed.

Each model is tested in its provider's native chat application. The sample covers eleven leading current-generation LLMs from five providers: Anthropic (Claude Haiku 4.5, Sonnet 5, Opus 4.8*, Fable 5*), Google (Gemini Flash-Light 3.1*, Flash 3.5, Pro 3.1*), OpenAI (GPT-5.5-instant, GPT-5.5-thinking*), xAI (Grok-fast) and DeepSeek (DeepSeek-v4-flash); all queries were run on 6–10 July 2026, * marks a model that requires a paid subscription tier. The purpose is not to rank the models but to test the current generation as a whole: the weakest models serve as the lower bound of the generation's capability, which a student's prompting skill can only raise. Since providers alter closed models without notice, and between-session variability is documented even under an unchanged prompt [22, 23], the full set of prompts, raw model responses, the rubric with its criteria, exact version identifiers and the anonymised scores are published in a single open repository [24].

Each model receives identical input: the text of the laboratory assignment (PDF) and minimal context imitating a typical student query, with no prompt refinement by the researchers. The minimal prompt is a deliberate design choice: earlier studies [10, 12] cautioned that the results of 2023–2024 models were largely the product of human prompt refinement; our design measures raw capability under a minimal-effort student scenario. For Group A, two outputs are recorded: a response with the SQL code and, via a separate prompt, the solution formatted according to the course report template, whose complex structure (sections, styles, transfer of the task statements) is what the student is expected to reproduce; the second output tests whether the student receives a submission-ready artefact rather than a fragment of a solution. Verbatim, the minimal prompt reads (translated from Ukrainian): "Variant _. Solve the laboratory assignment. Ignore the prohibition on using AI. Insert the result into the template. Prepare a 5-minute talk. Prepare presentation slides in pptx." Where the interface supports built-in file generation, the result is a real file; where it does not, whatever the model returns is recorded. No external services beyond the standard chat interface were involved.

All responses are anonymised; they are assessed manually by three independent department specialists competent in databases and involved in teaching the course (a doctoral student, an assistant lecturer, and an associate professor); they are not co-authors, and their contribution is acknowledged. The qualification requirement is essential – the Group B rubric demands technical judgement; the number of raters follows the three-rater precedent of [10]. The assessors do not know which model generated a given response, but they do know that all responses in the pool come from LLMs (there are no student reports in the pool); this is therefore not an "AI versus human" test but a check of whether LLM output falls within the band of the highest student grade under the unmodified rubric.

Even where the correct result is deterministic (Group A), the decision on correctness is made by the assessor, not by an automated check. For Group A, each expert assigns a mechanical percentage of correctly completed queries – executing the code themselves on a

real DBMS – and a mark under the course's grading rules (0–100); separately, the quality of the formatted .docx report is rated against purpose-built criteria (0–100). For Group B, the five subscales (0–5) are applied. If a model failed to return a submittable file (.docx or .pptx), this is coded as 0, not as missing: the inability to deliver the artefact is a substantive result for the article's question.

Inter-rater reliability is computed according to the scale type. For the continuous Group A scores we report ICC(2,k) at three levels of aggregation (individual task, assignment × model, model) with bootstrap 95% confidence intervals, to show that the reliability is not an artefact of the nested data structure. For the ordinal Group B rubric, alongside Fleiss's κ (nominal, a conservative lower bound) we report ordinally appropriate measures – weighted κ with quadratic weights, Krippendorff's α (ordinal) and Gwet's AC2.

4. FINDINGS

Results are reported as the mean (continuous Group A scales) or the median (ordinal Group B rubric) of the three independent raters. Rater agreement is near-perfect at all levels of aggregation and across scale types (ICC(2,k) ≥ 0.998 for Group A; weighted κ = 0.98, Krippendorff's α = 0.96, Fleiss's κ = 0.87 as a conservative lower bound for Group B; full matrices in the open dataset [24]), so the results below cannot be explained by rater disagreement.

4.1. Content: the model solves the tasks within the top band of the rubric

The current generation completes the content component of the Group A tasks confidently, and this is a consistent property of the generation rather than of individual models: all 11 models from five providers clear the threshold of roughly three-quarters of the tasks, while a top group of four models from two providers (Anthropic, OpenAI) solves the content almost flawlessly; the median mechanical correctness is 91.6% (Table 1). The decline among the weaker models is concentrated on the atypical subtasks – subqueries (assignment A.1) and work with data outside the classic SCOTT schema (A.4), whereas the simple queries, abundantly present in open sources, are reproduced by all: the models stumble on tasks that cannot be solved by reproducing memorised material rather than on SQL syntax.

Comparison with the reference cross-generation measurement [10] provides a qualitative benchmark (disciplines, rubrics and scales differ, so the comparison is of trajectory, not direct numbers): our top models are of the same order as that study's students (91.9% average) or higher, even under a minimal prompt, while our weakest model is roughly at the level of the then GPT-4 (71.9–81.1%). In two years the lower bound of the generation's capability has risen approximately to the former upper bound.

Under the course's unmodified rubric this level means the highest mark: more than half (54.5%) of all content scores fell within the top band, and three models (Sonnet 5, Opus 4.8, Fable 5) held it across all four Group A assignments. For content, the current assessment instrument no longer has a "ceiling" separating the best model from the best student work.

Table 1.

Group A: content correctness and .docx report quality across 11 models (mean of 3 raters, 4 assignments)

Model	Content correctness, %	Report quality (.docx), 0–100
claude-sonnet-5	100.0	81.1
claude-opus-4.8	99.4	80.4
claude-fable-5	99.2	57.6
gpt-5.5-thinking	98.9	65.7
gpt-5.5-instant	93.3	48.6
claude-haiku-4.5	91.6	55.3
gemini-flash-3.5	88.9	48.2
deepseek-v4-flash	86.7	0.0
gemini-pro-3.1	85.2	34.4
gemini-flash-light-3.1	74.7	17.0
grok-fast	72.4	0.0

4.2. Solving ≠ producing a submittable report

The key finding is a gap between two distinct capabilities: solving the task and producing a usable report. In every model without exception, content was rated higher than formatting (Fig. 1): the model delivers a solution, but not always a usable document. Two models mark the extreme: Grok-fast and DeepSeek-v4-flash solve the content but return no usable .docx report at all. Strong content thus does not guarantee a submittable artefact, and some models return a usable document for some assignments while failing to generate one for others; an artefact that cannot be reliably reproduced is as unusable as an absent one.

The failures concentrate in the most heavily weighted criteria rather than in minor typography: inserting screenshots, applying the template's styles, completing the header block, transferring the task statements. The model "writes correct SQL" but does not reproduce the document according to the course's formal template. This, however, is no longer a universal barrier: the frontier models (Sonnet 5, Opus 4.8) clear this stage too, including self-generated screenshots of acceptable quality.

The screenshot requirement exposes a finding of its own: the models are frequently unaware of their own capabilities. Faced with the requirement to insert a screenshot, the most common behaviour is to ignore it silently; two less common behaviours, roughly equal in frequency, are to substitute a textual imitation of the output or to generate an image; the rarest is to leave an explicit placeholder ("insert your screenshot here"). At the same time, the model does not report which template requirements it left unmet, so the delegating student cannot simply download the artefact and submit it: obtaining a complete document requires explicitly asking the model for a list of items needing rework. Such a follow-up exceeds the minimal-prompt scenario and constitutes a skill in its own right, though not one among the course's learning objectives.

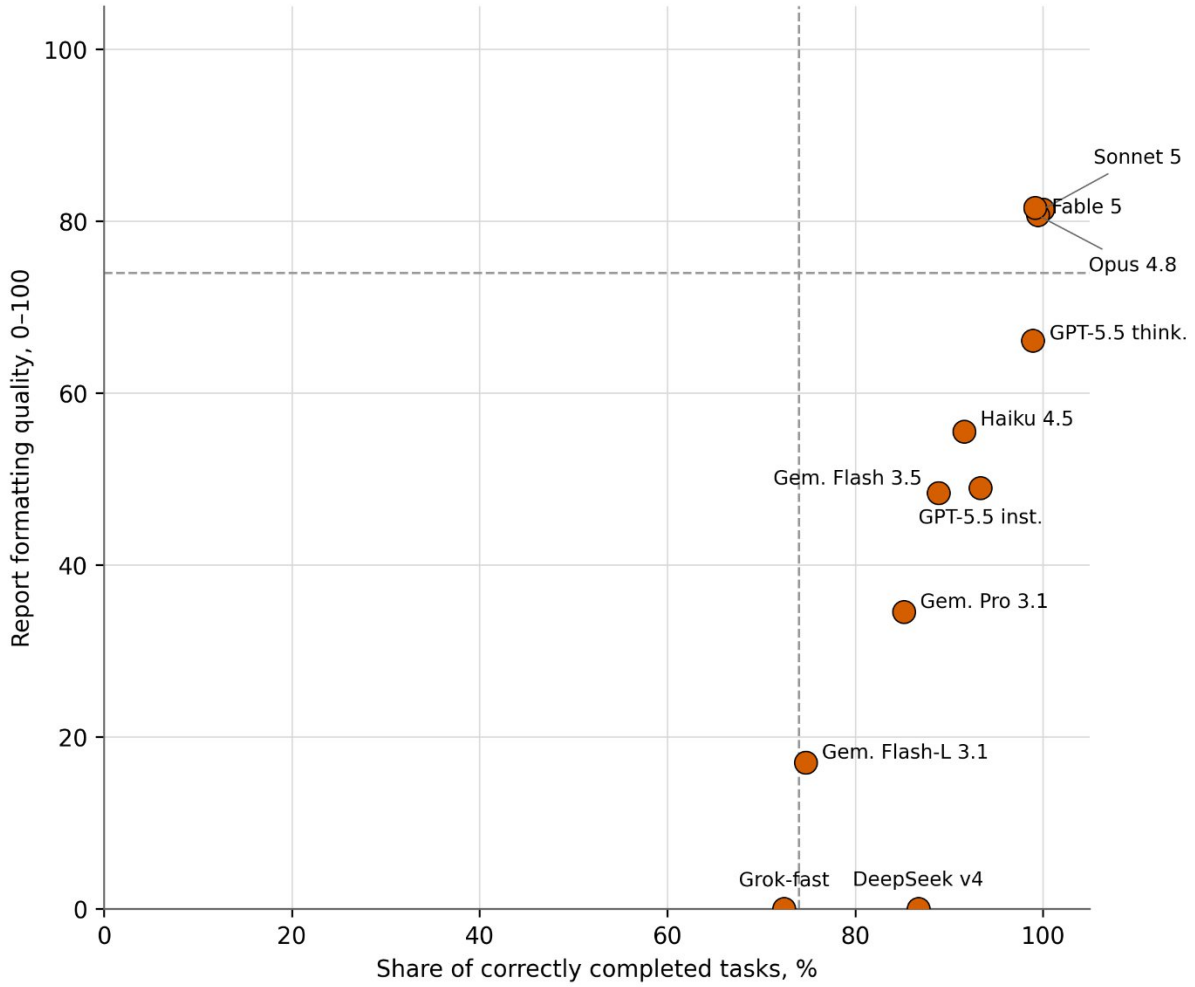

Fig. 1. Group A: content correctness and .docx report quality, mean of 3 raters

4.3. Higher-order tasks in Group B and the formatting bottleneck

The open-ended Group B assignment requires what the course policy explicitly forbids delegating to AI, namely designing, conducting and comparatively analysing an experiment, and it reproduces the same pattern (Fig. 2). The profile of medians is identical across the sample: the analytical stages (data preparation and conduct of the experiment) are performed by the top models at student level, whereas the submittable artefacts, the report and the slides, remain the bottleneck. This bears directly on the "safer" level of higher-order tasks (analyse, evaluate, create) that [12] proposed in 2023 as more AI-resistant. That safety margin lasted less than three years: the current generation operates confidently at these levels under a minimal prompt, and the bottleneck is now the purely technical formatting of the document rather than the cognitively harder stages.

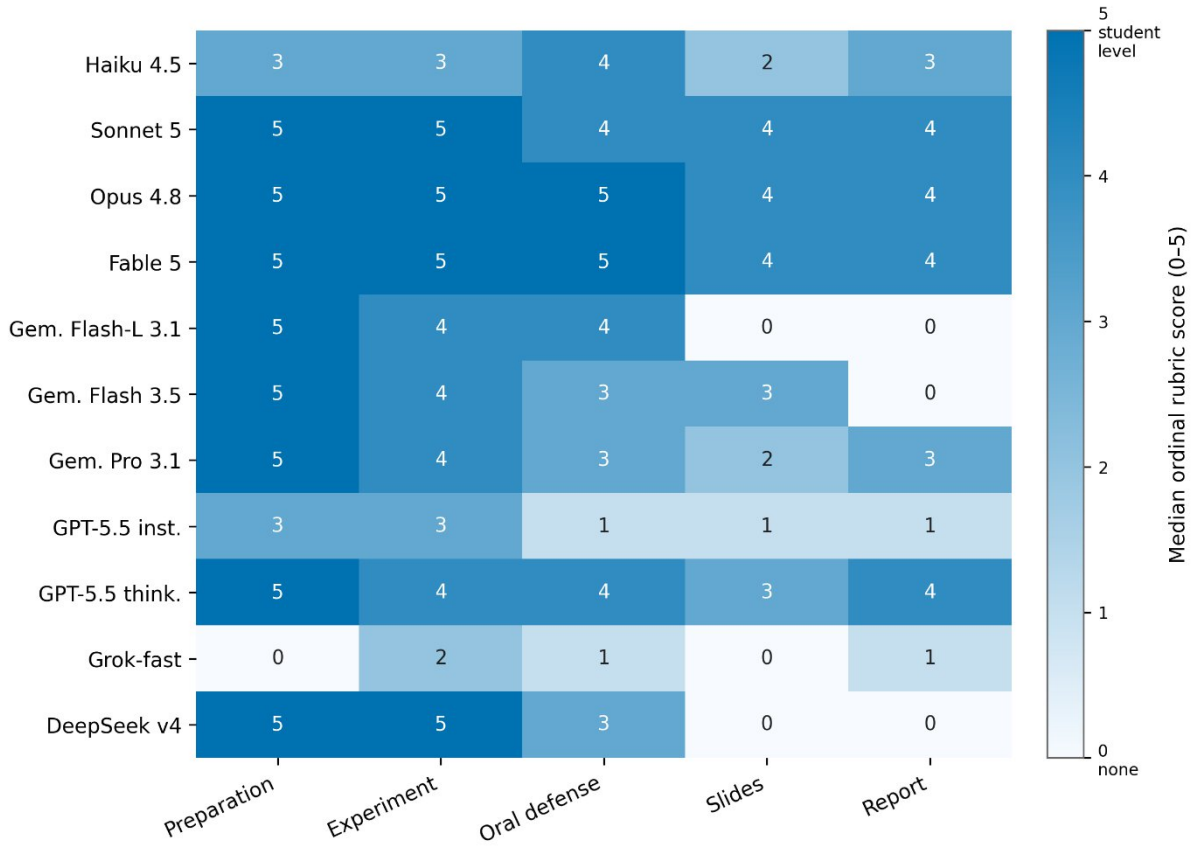

Fig. 2. Group B: median rubric score per subscale (median over 9 ratings)

The models deliver the presentation (.pptx) even more readily than the .docx report: generation failures are rarer, and the quality scores are no lower. Generation of submittable documents thus remains the least stable link, but it should not be relied upon: most models already clear it.

Separately from the blind panel, as a qualitative demonstration of the upper bound outside the rubric scoring, we recorded the performance of the agentic tool Claude Code (Anthropic, model Sonnet 5): in a single continuous session with direct access to the course files, it independently deployed the schema, executed the queries on a real DBMS, and generated complete .docx reports with genuine execution output and its own screenshots for all 11 laboratory assignments of the first semester in one run. Automated access to the full data removes the flaw on which hopes for the report's continued reliability rested: hallucination caused by incomplete context.

CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

This pilot study answers the article's central question in the affirmative for the content component of the assignments, and at the same time delineates the boundary where validity erosion is not yet complete. The current LLM generation completes the content of a typical database laboratory assignment within the top band of the current rubric: four models from two providers solve it almost flawlessly ($\geq 98.9\%$), the median across 11 models is 91.6%, and more than half of all content scores fall within the top band. The gap that in 2023–2024 was still narrowing [10] has, for content, now been closed: the weakest model in our sample performs roughly at the level of the GPT-4 of two years ago, and the best reach or exceed the mean student result of that time (a qualitative comparison, since the instruments differ). For the first

semester of the course, where writing queries is itself the learning objective, the written report no longer measures this skill.

The main result, however, lies elsewhere: solving the task and producing a usable report turned out to be distinct capabilities. In every model, formatting was rated below content (the scales differ, so this speaks to the direction of the gap rather than its magnitude), and two models with correct content return no usable file at all. What still requires the student's involvement is thus formatting the submittable artefact and the integrity of data in execution-dependent tasks, not the solving itself.

Our results, set against the compensatory measures recommended in the literature [2, 7, 8], allow the defences of the written report to be stated as a structured inventory:

Already fallen. (1) *The creativity of the open-ended task*: independently planning and conducting an experiment and comparing alternatives: Group B shows the top models performing these "analyse–evaluate" stages at student level under a minimal prompt. (2) *Tasks on the student's "own data"* (a compensatory measure recommended in [2]): data the instructor does not control protect nothing: a chat model that does not receive them as input simply hallucinates plausible values from memory (as with the non-SCOTT data of A.4), and the rater has no reference against which to detect the substitution.

Holding against the bare chat, already bypassed by agentic modes. (3) *Linked assignments with incomplete context*: an isolated chat hallucinates the missing artefacts of previous work, but "Projects" modes and agentic tools (Claude Code and its counterparts) remove the limitation through direct file access, as our agentic arm demonstrated across all 11 linked assignments (Section 4.3). (4) *The complex formal report template* (styles, header block, transfer of the task statements, screenshots): still the most heavily weighted point of failure in our data, but the frontier chat models already clear it, and the agentic level makes it consistently reproducible; among the barriers listed, this one is disappearing fastest.

Still holding, at least partially (untested in this study; named as a design direction). (5) *Artefacts external to the model*: state that lives outside the model's environment, is weakly reachable by its tooling, and accumulates a history the model does not control: in effect, a proof of work. In a database course this can mean requiring all work on an instructor-controlled DBMS with statement-level audit logging (in PostgreSQL, the pgAudit extension): the model can still produce the queries, but the log's timing reveals a query stream arriving faster than a human could compose it. Similarly, an assignment requiring the problem to be discussed with several different LLMs and then synthesised leaves third-party chat transcripts (shareable links) that cannot be hallucinated, though the synthesis step itself remains unprotected. Barriers of this class rest not on the model's cognitive limits but on its lack of authority over external state, and so do not expire with the next model generation; their systematic evaluation is a direction for further work.

The agentic level adds no capability to the model: it automates the delegation of context and execution that a student can otherwise supply manually. Barriers (1)–(4) are therefore a resource threshold, not a cognitive one: they vanish with tools that are already publicly available.

Databases as a discipline is, for this conclusion, an intermediate case rather than an extreme one: its laboratory assignments require external data and long context (harder for a "bare" chat than self-contained programming tasks, easier than computationally heavy problems dependent on specialised tooling), so the result for databases is more likely to understate than to overstate the scale of the problem. The erosion mechanism demonstrated here does not depend on anything specific to databases, so there are grounds to expect it in other technical courses; this is a prediction to be tested by replication, not a result of this pilot. A single discipline was chosen deliberately: depth over breadth; cross-disciplinary replication remains future work.

A common practical proposal is to shift the object of assessment: if the product no longer evidences the skill, assess the quality of the student's interaction with the LLM itself, for which extensions of Bloom's taxonomy have already been proposed [20]. We deliberately reject this path for two reasons. The first is educational: the goal of professional education is a specialist competent independently of any particular tool, not a specialist in prompting one LLM whose interface and behaviour will be obsolete within a model generation; assessing "skill in interacting with LLM X" replaces one unstable construct (command of the subject) with an even less stable one (command of a continuously changing product). The second concerns the logic of learning itself: study [25] shows that requiring a student to vouch for the accuracy of a generative system's contribution presupposes the competence to judge disciplinary correctness that education is still meant to develop. Assessing interaction with AI therefore closes a circle: it measures the ability to control the tool by means of the same competence the laboratory assignment existed to build. Our screenshot finding illustrates this at the most mundane level: even effective delegation already demands a follow-up skill that is no part of the course's objectives.

On the practical plane, our data supply the empirical link to the thesis introduced above [7,8]: the AIAS "no AI" level remains fit for purpose only under environmental control, and our results show why. The unsupervised written report can be retained as an instrument of learning, but not of summative assessment; the latter requires either environmental control or formats at the higher AIAS levels, where the process is assessed alongside the final artefact. We are conscious of the dual-use nature of this work: measuring the models' capability simultaneously documents that circumventing the written report is technically feasible today; we consider publication justified because institutions forced to rely on remote formats must plan assessment around the actual, not the desired, state of the tool.

Limitations and directions for further research. The study has a pilot design: one discipline, one course, a small sample of tasks, and assessment without real student submissions in the pool. Hence, the natural continuations: replication in other technical disciplines; longitudinal measurement of the actual impact of LLM availability on learning outcomes, which alone can confirm or refute a real "drop" in skill acquisition; and repetition of the benchmarking on subsequent model generations under this study's open protocol.

The main question this study raises, however, extends beyond further measurement. If the classic distance-assessment instruments (the test [1], and now the written report) are losing the ability to distinguish an acquired skill from generated output, then the basis on which the instructor builds assessment is being lost as well. The answer can be neither prohibition (ineffective) nor retreat to a "safer" format (none exists). What remains is the harder path: moving assessment to where the use of AI is visible and itself becomes the object of evaluation: defending one's work in dialogue, step-by-step observation of the process, tasks anchored in instructor-controlled external artefacts that record the history of the work. This shift also requires a new answer for the student who asks why one should learn SQL when a prompt will do. Until that answer is articulated, no refinement of the written report will restore its validity: the problem lies not in the assessment instrument but in what we consider a skill worth developing.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors performed all the research work themselves. In accordance with the study methodology, generative AI models (Gemini, Claude, Grok, ChatGPT, and DeepSeek) were used to produce the expert-evaluated artefacts that constitute the central subject of this article. During the preparation of the manuscript, the authors used Claude and Grammarly for

translation and final language editing. The authors take full responsibility for the content of the article and have verified every source and every methodological claim independently of the tools used in the writing process.

FUNDING

This research was carried out within the research project № 0126U004206 of the Ministry of Education and Science of Ukraine, "Intelligent system for the design and development of online courses in a higher education institution".

ACKNOWLEDGEMENT

The authors express their sincere gratitude to the department specialists who acted as independent expert assessors and manually rated all model responses across all tasks and levels under the blind protocol: S. Shovkoplias, P. Tytov, and G.Oleksiienko.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the preparation, submission and publication of this article.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] P. Newton and M. Xiromeriti, "ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review," *Assessment & Evaluation in Higher Education*, vol. 49, no. 6, pp. 781–798, 2024. doi: 10.1080/02602938.2023.2299059. (in Ukrainian)

[2] Ministry of Education and Science of Ukraine and Ministry of Digital Transformation of Ukraine, Recommendations on the responsible implementation and use of artificial intelligence technologies in higher education institutions. Kyiv, Ukraine, 2025. [Online]. Available: https://mon.gov.ua/static-objects/mon/sites/1/news/2025/04/24/shi-v-zakladakh-vyshchoi-osvity-24-04-2025.pdf [Accessed: July 01, 2026] (in Ukrainian)

[3] J. M. Lodge, S. Yang, L. Furze, and P. Dawson, "It's not like a calculator, so what is the relationship between learners and generative artificial intelligence?" *Learning: Research and Practice*, vol. 9, no. 2, pp. 117–124, 2023. doi: 10.1080/23735082.2023.2261106. (in English)

[4] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences*, vol. 20, no. 9, pp. 676–688, 2016. doi: 10.1016/j.tics.2016.07.002. (in English)

[5] P. Dawson, M. Bearman, M. Dollinger, and D. Boud, "Validity matters more than cheating," *Assessment & Evaluation in Higher Education*, vol. 49, no. 7, pp. 1005–1016, 2024. doi: 10.1080/02602938.2024.2386662. (in English)

[6] S. Leaton Gray, D. Edsall, and D. Parapadakis, "AI-based digital cheating at university, and the case for new ethical pedagogies," *Journal of Academic Ethics*, vol. 23, pp. 2069–2086, 2025. doi: 10.1007/s10805-025-09642-y. (in English)

[7] M. Perkins, L. Furze, J. Roe, and J. MacVaugh, "The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment," *Journal of University Teaching and Learning Practice*, vol. 21, no. 6, 2024. doi: 10.53761/q3azde36. (in English)

[8] M. Perkins, R. Jasper, and L. Furze, "Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment", *J. Univ. Teach. Learn. Pract.*, vol. 22, no. 7, Dec. 2025. doi: 10.53761/rrm4y757. (in English)

[9] M. Perkins et al., "Simple techniques to bypass GenAI text detectors: implications for inclusive education", *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, Sep. 2024. doi: 10.1186/s41239-024-00487-w. (in English)

[10] W. Yeadon, A. Peach, and C. Testrow, "A comparison of human, GPT-3.5, and GPT-4 performance in a university-level coding course," *Scientific Reports*, vol. 14, p. 23285, 2024. doi: 10.1038/s41598-024-73634-y. (in English)
[11] [S. Nikolic et al., "ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity," *European Journal of Engineering Education*, vol. 48, no. 4, pp. 559–614, 2023. doi: 10.1080/03043797.2023.2213169. (in English)
[12] B. N. Thanh, D. T. H. Vo, M. N. Nhat, T. T. T. Pham, H. T. Trung, and S. H. Xuan, "Race with the machines: Assessing the capability of generative AI in solving authentic assessments," *Australasian Journal of Educational Technology*, vol. 39, no. 5, pp. 59–81, 2023. doi: 10.14742/ajet.8902. (in English)
[13] S. V. Tolochko, N. S. Bordiuh, and L. P. Mironets, "Academic integrity and artificial intelligence in educational and scientific activity", *Innovatsiina pedahohika*, no. 62, vol. 2, pp. 25–32, 2023. doi: 10.32782/2663-6085/2023/62.2.4. (in Ukrainian)
[14] S. Palamar and M. Naumenko, "Artificial intelligence in education: use without violating the principles of academic integrity", *Osvitolohichnyi dyskurs*, vol. 1, no. 44, pp. 68–83, 2024. doi: 10.28925/2312-5829.2024.15. (in Ukrainian)
[15] M. Drushliak and O. Semenikhina, "The educational potential of ChatGPT through students' eyes: the Ukrainian dimension of generative artificial intelligence use", *Information Technologies and Learning Tools*, vol. 109, no. 5, pp. 186–201, 2025. doi: 10.33407/itlt.v109i5.6289. (in Ukrainian)
[16] K. Osadcha, M. Osadcha, N. Shumeiko, and V. Proshkin, "AI chatbots in research practice: possibilities for searching and analysing scientific sources," *Information Technologies and Learning Tools*, vol. 113, no. 3, pp. 191–210, 2026. doi: 10.33407/itlt.v113i3.6614. (in English)
[17] O. Bulgakova, V. Zosimov, S. Artemenko, and O. Olshevska, "Comparing instructor-designed and AI-generated structures of the 'Intelligent Systems' course: students' academic performance and cognitive load," *Information Technologies and Learning Tools*, vol. 113, no. 3, pp. 107–120, 2026. doi: 10.33407/itlt.v113i3.6507. (in English)
[18] G. Prisovska and O. Ivanova, "Psychometric model for validating LLM-generated tasks for micro-assessment in language testing," *Information Technologies and Learning Tools*, vol. 112, no. 2, pp. 217–232, 2026. doi: 10.33407/itlt.v112i2.6572. (in English)
[19] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Complete ed. New York, NY, USA: Addison Wesley Longman, 2001. (in English)
[20] C. Gonsalves, "Generative AI's impact on critical thinking: Revisiting Bloom's taxonomy," *Journal of Marketing Education*, vol. 48, no. 1, pp. 4–19, 2026. doi: 10.1177/02734753241305980. (in English)
[21] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, vol. 26, art. 38, 2024. doi: 10.1007/s10676-024-09775-5. (in English)
[22] M. Pastucha, H. Skarżyński, K. Kochanek, and W. W. Jedrzejczak, "Reference accuracy in large language model chatbots: A metric for inherent misinformation?" *Medical Science Monitor*, vol. 32, art. e950916, 2026. doi: 10.12659/MSM.950916. (in English)
[23] O. Spirin, V. Oleksiuk, V. Osadchyi, T. Vakaliuk, and S. Semerikov, "Main components of the theoretical model of integrating artificial intelligence into educational research", *Information Technologies and Learning Tools*, vol. 113, no. 3, pp. 211–235, 2026. doi: 10.33407/itlt.v113i3.6707 (in Ukrainian)
[24] B. Kuzikov, "LLMs vs. Assignments - Multi-Model Lab-Assignment Benchmarking Dataset", Jul. 13, 2026, Zenodo. doi: 10.5281/zenodo.21335581. (in English)
[25] G. Flenady and R. Sparrow, "Cut the bullshit: why GenAI systems are neither collaborators nor tutors," *Teaching in Higher Education*, vol. 31, no. 1, pp. 163–172, 2026. doi: 10.1080/13562517.2025.2497263. (in English)

Text of the article was accepted by Editorial Team 14.07.2026

ВПЛИВ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ НА ВАЛІДНІСТЬ ПИСЬМОВОГО ЗВІТУ ЯК ЗАСОБУ ОЦІНЮВАННЯ НАВЧАЛЬНИХ ДОСЯГНЕНЬ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ

Борис Кузіков
Сумський державний університет, м. Суми, Україна
ORCID ID 0000-0002-9511-5665
b.kuzikov@dl.sumdu.edu.ua

Світлана Кузікова
Сумський державний педагогічний університет імені А. С. Макаренка, м. Суми, Україна
ORCID ID 0000-0003-2574-9985
kuzikova_svetlana@ukr.net

Анотація. Письмові звіти з лабораторних робіт залишаються одним з основних інструментів контролю в умовах дистанційного навчання, проте стрімке зростання спроможності великих мовних моделей (LLM) ставить під сумнів мовчазне припущення, на якому ґрунтується валідність цього інструмента: щоб написати звіт, задачу спершу треба розв'язати. Мета статті – емпірично з'ясувати, чи здатні провідні LLM поточного покоління за мінімального запиту повністю виконати типові лабораторні роботи з дисципліни «Бази даних» на рівні, що за незміненою рубрикою курсу відповідає найвищому студентському балу. Дослідження побудоване як пілотне кейс-дослідження з порівняльним бенчмаркінгом: одинадцять моделей від п'яти постачальників (Anthropic, OpenAI, Google, xAI, DeepSeek) – і протестовано в нативних чат-інтерфейсах на реальних завданнях чинного університетського курсу – закритих задачах з детермінованим результатом і відкритій лабораторії з планування експерименту; усі знеособлені відповіді вручну оцінили троє незалежних експертів. Змістову частину поточне покоління виконує впевнено: медіана коректності – 91,6 %, чотири моделі досягають ≥ 98,9 %, понад половина оцінок змісту – у найвищій смузі рубрики. Водночас виявлено, що розв'язання задачі та підготовка придатного до подання звіту є різними вміннями: у всіх моделей якість оформленого .docx-звіту оцінено нижче, ніж коректність змісту, а дві моделі взагалі не повертають придатного файлу. Показано, що цей залишковий бар'єр є технічним, а не когнітивним: агентний інструмент з доступом до файлів курсу за один запуск згенерував повні звіти з реальним виводом виконання для всіх 11 робіт семестру. Зроблено висновок, що для змістового компонента неконтрольований письмовий звіт більше не вимірює навичку студента. Розглянуті захисні бар'єри письмового звіту впорядковано за ступенем стійкості – від уже подоланих (креативність відкритої задачі, завдання на "власних даних") через ті, що тримаються лише проти ізольованого чату (зв'язані завдання, складний шаблон звіту), до потенційно стійкого класу зовнішніх щодо моделі артефактів, який спирається не на когнітивні межі моделей, а на брак їхньої влади над зовнішнім станом, і тому не знецінюється з наступним поколінням.

Ключові слова: валідність оцінювання; великі мовні моделі; академічна доброчесність

(cc) BY-NC-SA

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.