

Review

Evaluating Generative AI Assisted Lesson Planning for Secondary Chemistry: A Critical Review of Planning Quality and Teacher Judgment

Ruiyang Gong¹, Xinyue Luo¹ and Yuling Zhao^{1,*}
¹ Zhejiang Normal University, Jinhua, China

* Correspondence: Yuling Zhao, Zhejiang Normal University, Jinhua, China

Abstract: Generative artificial intelligence (GenAI) can produce chemistry lesson plans containing objectives, activities, resources, and assessments, but evaluating their educational contribution requires different kinds of evidence. This critical narrative review examines GenAI-assisted planning for secondary chemistry through three objects of evaluation: the written plan, teachers' professional reasoning and revisions, and students' learning. Seven recent chemistry and science education studies provide the principal contemporary evidence, interpreted alongside research on pedagogical content knowledge, chemical representations, and constructive alignment. The synthesis identifies useful planning support while showing why completed templates, teacher perceptions, and classroom outcomes require separate interpretation. Chemistry-specific review should examine the conditions attached to scientific claims, the connections among representations, the reasoning elicited by tasks, and the responses that trigger instructional adjustment. A worked illustration using electrolysis of aqueous copper sulfate demonstrates how these criteria translate into changes to objectives, diagnostic questions, and follow-up teaching decisions. The article proposes a planning-review procedure and a teacher-education assignment that make justified revisions visible through planning records and discussion. Future studies can connect these records with classroom enactment and student explanations while separating AI assistance from additional opportunities to revise. The review positions teachers' transformation of generated materials as the link between the availability of planning support and the quality of the learning opportunities eventually offered to students.

Keywords: generative artificial intelligence; secondary chemistry; lesson planning; teacher judgment; pedagogical content knowledge

Received: 19 August 2026

Revised: 01 October 2026

Accepted: 13 October 2026

Published: 20 October 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Generative artificial intelligence (GenAI) has made complete lesson plans readily available to teachers. A request specifying a topic, age group, and lesson duration can produce objectives, activities, questions, and an assessment within the same response. In secondary chemistry, the educational value of this output depends on decisions that extend beyond completing a familiar planning template. Teachers must connect observable changes with particle explanations, select the conditions under which a chemical claim holds, and anticipate how students will interpret symbols and diagrams. A plan can contain all the expected headings while leaving these decisions unresolved.

Early chemistry studies illustrate both the range of planning support and the importance of the teacher's contribution. Four general chemistry instructors used ChatGPT-4 to develop different aspects of lessons on historical atomic experiments, with the conversations reflecting their differing objectives and prior knowledge [1]. A comparison involving high-school electrolysis examined a human teaching sequence alongside sequences produced with GPT-3.5 and GPT-4 [2]. These studies address how

planning materials can be generated and examined. The next question concerns the evidence needed to judge their educational quality.

Three objects of evaluation recur across this literature: the written plan, the teacher's reasoning while working with the plan, and what students subsequently learn. Each requires a different form of evidence. Expert review can identify a misleading particle diagram. A revision history can show whether a teacher recognized and corrected its misleading features. Student explanations can reveal whether the revised diagram helped learners distinguish the intended concepts. Treating these as successive questions makes research findings more interpretable and gives teacher educators a clearer basis for assessing planning assignments.

This review examines GenAI-assisted lesson planning through these three objects of evaluation, with secondary chemistry as its instructional focus. It asks which features of a generated plan require chemistry-specific scrutiny, how teacher judgment becomes visible during revision, and how planning research can connect its findings to classroom learning. The synthesis develops a practical review procedure and illustrates its use through electrolysis of aqueous copper sulfate. Its central argument is that the educational contribution of GenAI depends on the quality of teachers' transformations of generated materials and on the evidence those transformations are designed to elicit from learners.

Recent reviews provide a wider context for these questions. Vega-Baudrit and Rivera Álvarez examine GenAI in university chemistry through the chemistry triplet and Biggs' 3P model [3]. The present review concentrates on secondary lesson planning and the relationship between a proposed teaching resource and the teacher's decisions about using it. This narrower focus makes the planning process itself available for scrutiny: how a suggestion becomes an activity, how the activity reveals understanding, and how that evidence changes the next instructional step.

2. Review Approach and Analytical Organization

The article uses a focused critical narrative review, following the principle that a literature review should align its selection and synthesis procedures with its research question [4]. Sources were located through publisher searches, DOI searches, and citation connections using combinations of chemistry education, generative AI, ChatGPT, lesson planning, pedagogical content knowledge, and teacher judgment. The review draws on publications available by 15 October 2026. Priority was given to studies examining chemistry planning tasks, generated teaching materials, or teachers' interactions with GenAI during preparation. Selected science education studies provide adjacent evidence where their planning processes clarify the chemistry discussion.

For each recent study, attention was given to the instructional context, the planning task, the model where reported, the evidence collected, and the outcome actually examined. Seven studies provide the principal contemporary evidence in Table 1. Their contributions range from examining laboratory proposals to analyzing teachers' dialogue, written plans, and reflections. Publisher texts or abstracts established the reported findings; DOI metadata established bibliographic details. Foundational work on teacher knowledge, chemical representation, and instructional alignment supplies the conceptual criteria used in the synthesis. These sources serve different analytical functions and are discussed at the point where each informs the argument.

Table 1. Principal Studies and the Planning Evidence They Contribute

Study and context	Task and evidence	Contribution to this review
Clark et al. [1], introductory university chemistry	Four instructors used ChatGPT-4 for historical atomic-experiment lessons; instructor accounts of the conversations and uses.	Shows how objectives and prior knowledge shape the instructional assistance sought.

Gonçalves Costa et al. [2], Brazilian high-school electrolysis	Comparison of a human sequence with GPT-3.5 and GPT-4 sequences under shared planning specifications.	Supports analysis of plan content and sequencing; the AI sequences were not classroom learning interventions.
Araújo and Saúde [5], Portuguese grade-11 laboratory curriculum	Prompt refinement and examination of proposals for five mandatory laboratory activities.	Connects scientific and pedagogical review with safety and sustainability.
Akaygun and Kilic [6], chemistry teacher education	Preservice teachers reflected on GenAI chemistry visuals.	Makes representational judgment visible; the reported object is teachers' reflections.
Peikos and Stavrou [7], primary science planning	Exploration of ChatGPT planning through pedagogical content knowledge.	Provides adjacent evidence for using subject-teaching knowledge to structure interaction.
Aydin Gunbatar et al. [8], middle-school acids and bases	Six participants with differing teaching experience; CoRe plans, chats, reflections, and focus group.	Distinguishes interaction profiles through the way suggestions are sought and incorporated.
Ndlovu et al. [9], science teacher education and equilibrium	Nineteen third-year preservice teachers produced traditional and AI-supported RfLD plans.	Separates gains in written reasoning dimensions from continuing weaknesses in adaptive triggers.

The synthesis groups evidence by the object evaluated: plan quality, professional reasoning, or learning. It then asks what connects one object to the next. A recommendation about teacher preparation requires evidence about the decisions teachers make, while a claim about student understanding requires learner evidence. The synthesis compares studies within their instructional settings and reported outcome types. The chemistry examples and review questions developed below are analytical illustrations derived from this synthesis. They provide concrete decisions through which the argument can be examined and used in planning practice.

Generative AI assisted source discovery, evidence organization, and drafting of this manuscript.

The studies examine different educational settings. University chemistry and primary science inform the analysis as adjacent contexts; findings are interpreted at their reported level.

3. Chemistry planning as a problem of representation and alignment

Chemistry lesson planning requires teachers to manage relationships among observations, models of entities that cannot be seen directly, and symbolic descriptions. Johnstone explained how the demands associated with these domains can make science difficult to learn [10]. Taber subsequently examined the nature of the chemistry triplet and its implications for interpreting chemical knowledge [11]. For planning review, the relevant question is how students will move between the domains during a particular task. An AI-generated sequence that places a photograph, particle sketch, and equation on successive slides still needs an explicit explanation of what learners should connect across them.

Different representations also perform different instructional functions. Ainsworth's framework identifies how multiple representations can complement information, constrain interpretation, and support deeper understanding [12]. An electrolysis circuit diagram and a particle sketch, for example, explain different parts of charge transport. The circuit diagram locates the power supply and external connections; the particle sketch can show ions moving through the solution. Requiring students to connect these

representations gives the teacher evidence about whether they distinguish the charge carriers in each region. The practical review question becomes whether the task asks learners to use the representations in a way that serves the intended explanation.

Representational review also involves deciding which information to withhold until it becomes useful. A complete electrolysis diagram can contain the power supply, electrode names, polarity signs, ion movements, electron arrows, and both half-reactions. Presenting these features simultaneously can make it difficult to identify the relationship students are meant to explain. A teacher can begin with the circuit and observations, ask students to propose the electrode processes, and then add the particle representation to examine charge transport. This sequencing applies the functional view of multiple representations described by Ainsworth: each addition should resolve an interpretive problem created by the preceding task [12]. The review can therefore examine when a representation enters the lesson as well as whether it is chemically correct.

Expertise affects how these relationships are perceived. Kozma and Russell documented differences between experts' and novices' responses to representations of chemical phenomena [13]. A teacher who finds a diagram immediately meaningful can therefore overlook the interpretive work it requires of a beginner. GenAI increases the number of candidate explanations and visuals available, making this judgment more frequent. The teacher needs to specify which features carry chemical meaning, which are decorative, and what prior knowledge the student needs to interpret the resource. A visually coherent image can obscure these questions when its attractiveness becomes the main reason for selecting it.

Pedagogical content knowledge (PCK) provides a basis for judging such choices. Shulman connected subject understanding with the forms through which content becomes teachable [14]. Technological pedagogical content knowledge (TPACK) extends attention to the relationships between content, pedagogy, and technology [15]. Applying these ideas to GenAI-assisted planning means evaluating a proposed explanation through a specific learner difficulty and a specific instructional purpose. Constructive alignment supplies a further criterion: objectives, activities, and assessment should require the same kind of intellectual performance [16]. An objective about explaining charge transport requires more than naming the electrodes.

Alignment can be inspected at the level of a single question. Suppose the objective asks students to explain why two observations are consistent with one chemical model. A task asking them to label apparatus provides preparation for that explanation but does not elicit it. The plan needs a subsequent question requiring students to make and justify the connection. This distinction helps a reviewer decide which activity is essential, which provides support, and which consumes time without serving the stated objective. GenAI can offer alternative tasks once the missing intellectual performance has been identified, giving the interaction a definite instructional target.

4. What the recent evidence establishes

The contemporary studies show that GenAI can supply several components of a plan, while the meaning of improvement depends on the task being evaluated. Clark et al. reported uses including lesson outlines, resources, explanations, and assessment questions in university chemistry [1]. In high-school electrolysis, Gonçalves Costa et al. compared sequences developed under common specifications, including lesson number and duration [2]. Holding these specifications constant makes differences in the proposed instructional content easier to examine. Their comparison concerns the teaching sequences as planning artifacts; the human sequence had a classroom history that the two AI-generated comparators did not share.

Laboratory planning introduces decisions with direct practical consequences. Araújo and Saúde examined ChatGPT proposals against the Portuguese grade-11 laboratory curriculum and highlighted scientific-pedagogical suitability, safety, and sustainability [5]. Their work directs attention to details that a general lesson template can leave implicit: what substances are used, what operations students perform, and how materials are

handled after the activity. These details also shape the learning task. A procedure that requires students merely to follow instructions elicits different evidence from one requiring them to justify a measurement choice or explain an unexpected observation.

Studies of teachers' interactions reveal variation that a finished plan conceals. Aydin Gunbatar et al. examined six participants preparing middle-school acids-and-bases plans, using CoRe documents, chat histories, reflections, and a focus group [8]. They identified AI-reliant, AI-collaborative, AI-assisted refinement, and AI-independent planning profiles. The profiles describe different ways of seeking and incorporating suggestions. This supports examining who makes consequential decisions within the planning process. Two teachers can submit similarly organized plans while differing in whether they have assessed the proposed misconception, selected a suitable representation, or accepted the chatbot's recommendation unchanged.

Visual-resource research makes the object of evaluation especially clear. Akaygun and Kilic analyzed chemistry preservice teachers' reflections on GenAI visuals, including the aspects participants regarded as strengths and limitations [6]. The evidence concerns the participants' interpretations and decisions. Evaluating a generated image's scientific accuracy requires an additional comparison with explicit chemical criteria. This distinction leads to a useful teacher-education task: ask a student teacher to identify the meaning of each visual feature and support the judgment with a relevant chemical model. The task records more than whether the participant liked or disliked the image.

The equilibrium study by Ndlovu et al. brings the review close to secondary chemistry teacher preparation [9]. Nineteen third-year science preservice teachers first produced traditional plans and then AI-supported RfLD revisions. Rubric-based analysis found stronger written anticipation of student thinking, instructional justification, and alignment, while adaptation remained in the same limited performance category. The revised plans still lacked clear triggers for changing instruction. Because the tasks occurred in a fixed sequence, the comparison combines AI support with another opportunity to plan; the study also did not record the ChatGPT version. These design features identify concrete controls for further research on planning improvement.

Taken together, the studies direct attention to the different demands placed on a planning tool. Producing a recognizable teaching sequence requires organization of content and activities. Adapting that sequence requires a judgment about which feature of a learner's response should change instruction. An evaluation that rewards the presence of differentiated activities can register improvement even when the plan gives no reason to select one activity over another. This helps explain the significance of adaptive triggers in the synthesis. The planning record becomes more informative when it connects a stated difficulty to a diagnostic response and then to a particular instructional action, allowing a reviewer to inspect the logic linking them.

5. Evaluating the quality of a generated chemistry plan

The first stage of review is to make the chemical claim and its conditions explicit. For an electrolysis lesson, the proposed products depend on the electrolyte, its concentration, the electrodes, and the operating conditions. A plan that requests the products without specifying the cell has omitted information needed to judge a student's answer. The teacher should complete these conditions before deciding whether a question is suitable. The same principle applies to equilibrium: a statement about changing concentration needs a defined reaction, a stated temperature where relevant, and a distinction between the immediate disturbance and the subsequent response. This is an efficient review step because it often exposes the source of an otherwise ambiguous explanation.

A useful revision record separates the location of a problem from the reason for changing it. In a generated electrolysis plan, the problem might appear in an answer key, but originate in an unspecified electrode material. Correcting the answer alone leaves the task unstable because another reader could reasonably assume a different cell. The teacher should revise the task conditions and then check the dependent diagram and marking criteria. The laboratory proposals examined by Araújo and Saúde make the suitability of

generated instructions a concrete object of review [5]. Applied to lesson preparation, this approach asks the teacher to trace how one chemical condition affects several parts of the planned learning sequence.

Representation review then examines the relationship between the chosen medium and the targeted reasoning. If a particle diagram shows equal numbers of reactant and product particles at equilibrium, students need to know whether this is one illustrative composition or a general claim. A follow-up question can ask whether unequal amounts would rule out equilibrium and what further information would be required. The teacher should also check particle identity, conservation, and the meaning of arrows or colors. Applying these criteria gives a chemical reason for retaining, editing, or replacing a visual. It also converts resource selection into a decision that can be discussed with a mentor.

A further check concerns the learner's opportunity to construct an explanation. A sequence can name inquiry, collaboration, or discussion while assigning tasks that require only copying statements. The review should identify exactly what students predict, compare, measure, or justify. For example, placing a prediction before a simulation and requesting a revised explanation afterward gives the teacher two interpretable pieces of learner work. Their difference can guide feedback. The time needed to produce and discuss that work should appear in the plan, so the assessment is part of the instructional sequence rather than an activity added after the available lesson time has been used.

Differentiation can also be reviewed through a shared chemical objective. One group could receive a circuit with selected labels already supplied, while another examines an explanation containing a charge-carrier error. Both groups can finish by explaining the same relationship between ion movement and electrode reactions. The teacher can then compare responses against common criteria while adjusting the support students receive. This design gives the term differentiation a specific instructional meaning. A generated plan that assigns a simpler recall task to one group needs revision when that task removes the explanation named in the objective. The review should identify which demand remains shared and which support is being varied.

Practical feasibility should be evaluated at the level of classroom actions. A laboratory task needs quantities, equipment, supervision arrangements, and an appropriate local procedure for handling materials. A simulation task needs an accessible resource and a question that makes use of its adjustable variables. If students are expected to compare outputs, the plan should specify what they record and how they organize that record. These details allow the teacher to decide whether the proposed activity is achievable with the available resources. They also reveal whether replacing a practical activity with another medium preserves the intended scientific reasoning.

Assessment quality depends on the inference a response will support. A multiple-choice item about equilibrium can identify a selected statement, but a short explanation reveals why the student selected it. An effective planning review therefore pairs an expected response with at least one plausible alternative and specifies what the teacher would do next. Table 2 assembles these questions into a review procedure. Each row produces a visible revision decision. The procedure can be applied to a whole plan or to one consequential component, such as an experimental instruction, a diagram, or an exit question.

Table 2. Chemistry-specific Questions for Reviewing a GenAI-assisted Plan

Review focus	Question for the teacher	Revision evidence
Chemical claims	Are the reaction, materials, and conditions sufficient to determine the explanation or answer?	A corrected claim with the conditions stated.
Representations	What does each visible feature mean chemically, and what connection must students make?	An annotated diagram, graph, or equation with a linking question.

Learner reasoning	What prediction or explanation will make students' thinking accessible?	A task and a planned record of student responses.
Feasibility and laboratory practice	Can the activity be carried out with the stated time, equipment, and approved procedures?	A completed resource specification or a reasoned replacement activity.
Alignment	Does the assessment require the reasoning named in the objective?	An assessment item and a criterion tied to the objective.
Adaptive decisions	Which response would trigger a change of explanation, representation, or task?	An explicit decision point with a corresponding teaching response.

The order of these checks matters for the amount of revision required. Chemical claims and task conditions should be examined before refining the language or appearance of the lesson. An incorrect reaction model can affect the objective, diagram, worked answer, and assessment simultaneously. Correcting it early prevents the teacher from polishing several dependent components that will later need replacement. Once the chemistry and alignment are coherent, attention can shift to wording, accessibility, and pacing. This order makes the procedure manageable for novice teachers because it directs their limited planning time toward decisions with the largest instructional consequences.

The questions are a proposed planning-review procedure developed in this article. Teachers apply them using the curriculum and laboratory requirements of their own setting.

6. A worked planning illustration electrolysis of aqueous copper sulfate

Electrolysis of aqueous copper sulfate illustrates how a general plan can be transformed into a sequence of assessable decisions. The instructional model specifies ideal inert electrodes, a direct-current supply, and copper(II) ions remaining available throughout the modeled interval. Students receive a supplied circuit diagram and an observation record describing copper deposition at the cathode and gas evolution at the anode. The objective is to explain these changes through electrode reactions and charge transport. Using supplied materials directs the planning review toward the chemical reasoning required to connect the observations, particle model, and circuit.

The teacher checks the proposed half-reactions against the stated conditions. Cathodic deposition is $\text{Cu}^{2+}(\text{aq}) + 2\text{e}^- \rightarrow \text{Cu}(\text{s})$, while anodic water oxidation is $2\text{H}_2\text{O}(\text{l}) \rightarrow \text{O}_2(\text{g}) + 4\text{H}^+(\text{aq}) + 4\text{e}^-$. Doubling the cathodic equation balances the electron transfer when the half-reactions are combined. Students then identify where electrons occur in the equations and how charge moves between the electrodes through ionic migration in solution. The diagram should distinguish these regions: electrons are transferred through the external metallic circuit, and ions carry current within the electrolyte. Each arrow needs a defined carrier and a location consistent with that distinction.

A short counting task can make the electron relationship accessible before introducing a quantitative current calculation. Suppose the supplied particle model records the reduction of four copper(II) ions. Students must account for eight electrons at the cathode and use the anodic equation to identify the corresponding oxidation of four water molecules, producing two oxygen molecules and eight hydrogen ions. They can check atoms and charge in the combined process. The matching electron counts connect the electrode processes through the external circuit. This task gives the teacher a way to distinguish an error in balancing from an error in locating charge carriers, even when both students have reproduced the same memorized half-reactions.

The accompanying particle sketch should represent the electrolyte as a solution containing mobile ions. Sulfate ions remain part of the charge-transport explanation even though they do not appear in the chosen electrode half-reactions. Asking why they are absent from those equations opens a focused discussion of the difference between migration and reaction at an electrode. The teacher can compare a sketch that omits these ions with one that includes them and ask which better explains conduction in the solution.

If a generated image depicts the aqueous electrolyte as isolated neutral copper sulfate units, this comparison supplies a chemical reason to replace the image.

The first diagnostic question asks students to complete and explain charge-carrier arrows on the diagram. If a response sends electrons across the solution from one electrode to the other, the teacher returns to the particle sketch and asks which charged species are present in the aqueous region. The revised explanation must connect ion movement with the electrode processes. A second question asks students to identify the cathode from its half-reaction and then determine its connection to the supply. If a student identifies the cathode as positive by recalling a galvanic-cell diagram, the teacher uses electron gain to locate reduction and traces the connection to the negative terminal in this electrolytic cell.

GenAI can supply candidate questions once these diagnostic targets are specified. A focused request can provide the cell conditions, require separate assessment of electron versus ion transport and cathode polarity, and ask for reasoning behind each answer option. The teacher checks the proposed answers against the circuit and half-reactions before selecting an item. Table 3 records examples of this transformation. The planned follow-up depends on the student's response: a transport error requires revisiting charge carriers, whereas a polarity error requires relating reduction to the external supply. The resulting plan contains explicit triggers for instructional adjustment.

Table 3. Illustrative Revisions for an Aqueous Copper Sulfate Electrolysis Lesson

Planning component	Hypothetical initial wording	Proposed revision and its purpose
Objective	Understand electrolysis.	Explain copper deposition, anodic water oxidation, and charge transport for the stated inert-electrode cell.
Cell specification	Predict the products of electrolysis.	Provide aqueous CuSO_4 , ideal inert electrodes, the supply connections, and continued availability of Cu^{2+} before asking for products.
Representation task	Show the movement of electrons.	Label electron transfer in the external wires and ionic migration in solution; explain how they connect with the electrode reactions.
Adaptive response	Explain again if students are confused.	For electrons drawn across solution, revisit aqueous charge carriers; for a positive cathode claim, trace reduction to the supply's negative terminal.
Focused GenAI request	Make this lesson more engaging.	Generate separate diagnostic items for charge carriers and cathode polarity in the specified cell; explain each answer using the half-reactions and circuit.

Constructed examples illustrating the proposed planning-review procedure.

The reviewed lesson can retain a brief written decision record beside its main sequence. For the transport question, the record identifies the response that places electrons in solution, the particle sketch used for follow-up, and the explanation requested afterward. For the polarity question, it identifies the positive-cathode response and the circuit connection used to reconsider it. A teacher reviewing the plan can judge whether each response addresses the diagnosed difficulty. These records also guide a focused GenAI request: the teacher can ask for another question that elicits the same reasoning while preserving the specified cell conditions, then compare the resulting alternatives before choosing one.

The exit task presents the same modeled cell with a changed diagram layout and asks students to explain the cathode connection, identify charge carriers in the solution and wires, and balance electron transfer between the half-reactions. These responses show

whether students can coordinate the representations when familiar visual positions change. The teacher can then decide whether the next lesson needs further work on circuit connections or on the particle explanation. This sequence carries the planning argument into concrete learner evidence while keeping the chemistry tied to the specified electrolyte, electrode material, and availability of copper(II) ions.

7. Making teacher judgment visible in teacher education

Teacher education can assess the reasoning behind revisions by asking student teachers to submit a small planning record. A useful record contains an initial outline, the relevant GenAI exchange, the revised teaching component, and an explanation of the change. The explanation should name the chemical issue, the intended learner response, and the evidence used to justify the choice. This gives the mentor access to decisions that disappear in a polished final plan. It also permits assessment of a student teacher who rejects a generated suggestion for a sound chemical or pedagogical reason.

PCK can organize this record around a topic-specific difficulty. Peikos and Stavrou provide adjacent primary-science evidence for examining ChatGPT planning through PCK, while the chemistry interaction profiles reported by Aydin Gunbatar et al. show why the process of incorporating suggestions deserves attention [7] [8]. In a secondary chemistry assignment, the mentor can ask the student teacher to defend one selected representation and one assessment question. A short oral explanation provides an opportunity to probe the reasoning, especially when the written plan contains an elaborate rationale. The student teacher must connect that rationale to the actual task students will perform.

An oral planning conference can examine whether a student teacher's reasoning survives a change in the task. The mentor might supply a second circuit diagram with the electrode positions reversed on the page and ask which parts of the explanation require revision. A student teacher should identify the power-supply connections and electrode reactions before deciding what to change. This small variation exposes dependence on a familiar visual arrangement and gives the mentor a clear basis for feedback. The conference can then return to the submitted plan and identify where a learner would encounter a similar difficulty, connecting the student teacher's own explanation with a decision about instruction.

Peer discussion can then compare the consequences of alternative revisions. Two groups might retain different electrolysis diagrams while pursuing the same learning objective. Each group explains which misconception its diagram addresses and what student response would reveal an interpretive problem. Their comparison can expose a neglected condition or an unsupported connection between representations. Where access to teaching practice exists, the revised component can be used in a short teaching episode and examined through learner work. This connects planning with the evidence needed for the next revision and develops a habit of asking what an instructional decision makes observable.

Assessment criteria should therefore reward a justified transformation, the chemical coherence of the resulting task, and the planned use of learner evidence. Prompt complexity by itself offers little information about any of these outcomes. A concise request followed by careful correction can produce stronger instructional reasoning than a lengthy request followed by unexamined acceptance. This assessment approach also accommodates uneven access to GenAI tools: the central evaluated work is the teacher's explanation of a pedagogical decision. A shared generated plan can provide the same starting material for all participants when tool access is limited.

Mentor feedback can remain equally focused. Instead of commenting on every sentence of a plan, the mentor selects a decision that affects the learning objective and asks the student teacher to explain its consequences. For example, changing an assessment item from definition recall to graph interpretation changes both the expected response and the feedback students will need. The student teacher can revise these linked components and explain why they belong together. Repeating this cycle across different chemistry topics

builds a record of how the teacher's reasoning develops and helps identify which decisions still depend heavily on external suggestions.

Connecting planning research with classroom evidence

Future studies can strengthen the link between plan quality and teacher judgment by preserving the sequence of revisions. Researchers can compare an initial plan, the generated suggestions, the teacher's selections, and the final plan using explicit chemistry criteria. This design identifies whether the teacher repaired an error, introduced an error, or retained an unresolved issue. Repeated planning tasks on different topics would also help distinguish a temporary improvement in one document from a developing ability to evaluate instructional materials. Recording the model, access date, prompt, and relevant output makes the source of a suggestion traceable across these comparisons.

Comparative studies need to separate AI assistance from opportunities to revise. The fixed ordering in Ndlovu et al. makes this a concrete design issue [9]. A comparison between an initial unaided plan and an AI-supported revision gives the second task both additional resources and additional planning time. Studies can provide a parallel unaided-revision condition, balance task order, or assign comparable topics across conditions. Evaluation should then report the dimensions separately. Chemical accuracy, alignment, and adaptive planning may respond differently to the same intervention, as the equilibrium study's rubric patterns suggest.

Classroom research can examine whether the reviewed decisions affect the thinking they target. For an electrolysis lesson, suitable evidence includes explanations connecting electrode reactions with charge transport and responses to a changed circuit diagram. Observation can establish whether the planned diagnostic question was asked and whether the teacher used the resulting responses to adjust instruction. Collecting these forms of evidence together reveals where the planning-to-learning connection holds or breaks down. It also directs improvement toward a specific component: a missed diagnostic opportunity, an ineffective explanation, or a representation that students interpreted differently from the teacher's expectation.

A study linking plans to learner responses also needs marking criteria that match the proposed explanation. For the electrolysis example, separate criteria can address the electrode reaction, the identity of the charge carrier, and the connection to the supply. A response may correctly identify copper deposition while still placing electrons in the electrolyte. Keeping these elements visible avoids treating partial success as complete understanding. Researchers can examine how an instructional revision affects the distribution of explanations across these elements. Examples of the responses used to assign each category would make that interpretation accessible to readers and support comparison across successive versions of a lesson.

The choice of comparison task should follow the same logic. Repeating the original diagram mainly examines whether students can reproduce the explanation in its familiar arrangement. Changing the layout while preserving the cell conditions examines whether they can use chemical and circuit relationships to reconstruct the explanation. A subsequent task that changes electrode material asks a further question because it can change the electrode process itself. Planning research should identify which relationship remains constant across its tasks and which new demand is introduced. This specification makes it possible to explain a change in student performance through the content of the task and the instruction actually provided.

Plan review and learner-response marking should also use clearly specified units of judgment. A reviewer may evaluate one sentence for chemical accuracy, a task for alignment, and a sequence for responsiveness. Assigning one overall score to all three makes it difficult to identify what improved after revision. Researchers can retain a profile across these units and illustrate disagreements with the relevant plan excerpts. Where reviewers disagree, the report can explain whether the dispute concerns an omitted condition, an interpretation of the learner task, or the adequacy of a proposed follow-up. This gives the reader a route from the judgment back to the material being judged and

makes a subsequent replication more precise about the instructional feature it needs to examine.

Research should also account for the effort needed to obtain the final teaching resource. Planning time includes generating suggestions, checking chemistry, changing activities, and preparing materials for the classroom. Recording these components would clarify where GenAI saves effort and where it redistributes work. A quick initial response followed by substantial correction has a different practical value from a targeted response that fits an established plan. Combining effort records with the quality of the final task would support a more useful comparison for schools and beginning teachers than reporting the speed of text generation alone.

Conclusion

Evaluating GenAI-assisted chemistry planning requires attention to the plan, the decisions that produced it, and the learning evidence it is designed to elicit. The reviewed studies establish useful forms of planning support and show variation in how teachers incorporate generated material. Their findings are most informative when read through the particular object each study examined. For secondary chemistry, the decisive review questions concern chemical conditions, relationships among representations, and the reasoning students will make visible through the proposed tasks.

A practical response is to organize planning around justified revision. Teachers can inspect a consequential chemical claim, connect the chosen resource to a learner difficulty, and specify a response that would trigger instructional adjustment. Teacher educators can assess these decisions through planning records and discussion. Research can follow the same decisions into classroom enactment and student explanations. This approach gives GenAI a defined role in producing alternatives while making the teacher's selection, transformation, and use of evidence available for serious educational evaluation.

References

1. T. M. Clark, M. Phaner, M. Stoltzfus, and M. S. Queen, "Using ChatGPT to support lesson planning for the historical experiments of Thomson, Millikan, and Rutherford," *Journal of Chemical Education*, vol. 101, no. 5, pp. 1992–1999, 2024, doi: 10.1021/acs.jchemed.4c00200.
2. G. Gonçalves Costa, W. J. D. Nascimento Júnior, M. N. Mombelli, and G. Giroto Júnior, "Revisiting a teaching sequence on the topic of electrolysis: A comparative study with the use of artificial intelligence," *Journal of Chemical Education*, vol. 101, no. 8, pp. 3255–3263, 2024, doi: 10.1021/acs.jchemed.4c00247.
3. J. R. Vega-Baudrit and A. Rivera Álvarez, "Generative artificial intelligence in university chemistry education: A critical review using Johnstone's chemistry triplet and Biggs' 3P model," *Frontiers in Education*, vol. 11, Art. no. 1918707, 2026, doi: 10.3389/educ.2026.1918707.
4. H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019, doi: 10.1016/j.jbusres.2019.07.039.
5. J. L. Araújo and I. Saúde, "Can ChatGPT enhance chemistry laboratory teaching? Using prompt engineering to enable AI in generating laboratory activities," *Journal of Chemical Education*, vol. 101, no. 5, pp. 1858–1864, 2024, doi: 10.1021/acs.jchemed.3c00745.
6. S. Akaygun and I. Kilic, "Generative artificial intelligence (GenAI) as the artist of chemistry visuals: Chemistry preservice teachers' reflections on visuals created by GenAI," *Journal of Chemical Education*, vol. 102, no. 7, pp. 2549–2564, 2025, doi: 10.1021/acs.jchemed.4c00775.
7. G. Peikos and D. Stavrou, "ChatGPT for science lesson planning: An exploratory study based on pedagogical content knowledge," *Education Sciences*, vol. 15, no. 3, Art. no. 338, 2025, doi: 10.3390/educsci15030338.
8. S. Aydin Gunbatar, G. Tezcan Sirin, O. C. Ilkyaz, and Y. Mutlu, "Exploring the artificial intelligence interaction profiles of participants with different levels of teaching experience for lesson planning in the context of acids and bases," *Chemistry Education Research and Practice*, vol. 26, no. 4, pp. 977–995, 2025, doi: 10.1039/D5RP00064E.
9. B. P. Ndlovu, H. C. Khoza, and D. Sibanda, "Pre-service science teachers' adaptive pedagogical reasoning in AI-supported lesson plans: A case with chemical equilibrium," *Frontiers in Education*, vol. 11, Art. no. 1811975, 2026, doi: 10.3389/educ.2026.1811975.
10. A. H. Johnstone, "Why is science difficult to learn? Things are seldom what they seem," *Journal of Computer Assisted Learning*, vol. 7, no. 2, pp. 75–83, 1991, doi: 10.1111/j.1365-2729.1991.tb00230.x.
11. K. S. Taber, "Revisiting the chemistry triplet: Drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education," *Chemistry Education Research and Practice*, vol. 14, no. 2, pp. 156–168, 2013, doi: 10.1039/C3RP00012E.

12. S. Ainsworth, "DeFT: A conceptual framework for considering learning with multiple representations," *Learning and Instruction*, vol. 16, no. 3, pp. 183–198, 2006, doi: 10.1016/j.learninstruc.2006.03.001.
13. R. B. Kozma and J. Russell, "Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena," *Journal of Research in Science Teaching*, vol. 34, no. 9, pp. 949–968, 1997, doi: 10.1002/(SICI)1098-2736(199711)34:9<949::AID-TEA7>3.0.CO;2-U.
14. L. S. Shulman, "Those who understand: Knowledge growth in teaching," *Educational Researcher*, vol. 15, no. 2, pp. 4–14, 1986, doi: 10.3102/0013189X015002004.
15. P. Mishra and M. J. Koehler, "Technological pedagogical content knowledge: A framework for teacher knowledge," *Teachers College Record*, vol. 108, no. 6, pp. 1017–1054, 2006, doi: 10.1111/j.1467-9620.2006.00684.x.
16. J. Biggs, "Enhancing teaching through constructive alignment," *Higher Education*, vol. 32, no. 3, pp. 347–364, 1996, doi: 10.1007/BF00138871.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.