



Received: 29-08-2026
Accepted: 09-10-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Register Accommodation in ChatGPT-Generated Primary Science Explanations: A Conceptual Framework and Research Protocol

Konstantinos T Kotsis

Department of Primary Education, University of Ioannina, Greece

Corresponding Author: Konstantinos T Kotsis

Abstract

Generative artificial intelligence can produce science explanations addressed to children, yet fluent prose and favourable readability scores do not establish that an explanation is linguistically appropriate or scientifically comprehensible. This conceptual and methodological article develops a framework for examining register accommodation in ChatGPT-generated primary science explanations. Drawing on systemic functional linguistics, corpus-based register research, and recent studies of generated educational texts, it distinguishes lexical substitution from coordinated changes in grammatical packaging, causal explicitness, technical vocabulary, and audience positioning. Lexical density is treated as a descriptive property rather than a direct measure of

comprehension. A proposed corpus study will compare explanations of thirty science topics across three audience conditions and repeated generations with a matched human-authored corpus. Complementary age-only and register-guided prompts, independent scientific review, manual validation of automated annotations, and multilevel analysis are specified. Constructed examples illustrate how apparently simple language can retain conceptual compression or introduce misleading agency. No corpus has been collected and no empirical findings are reported. The article offers a testable research programme and a basis for teacher evaluation, while reserving claims about learning benefits for subsequent investigations involving children.

Keywords: Generative Artificial Intelligence, Science Education, Register Accommodation, Lexical Density, Primary Education

1. Introduction

Learning science requires learning ways of using language to identify phenomena, relate observations, formulate explanations, and distinguish evidence from interpretation. The linguistic demands of a science text therefore extend beyond the unfamiliarity of individual words. A child may recognise every word in a sentence and nevertheless fail to recover the process, relationship, or assumption that the sentence expresses. In the systemic functional account developed by Halliday and Martin^[1], scientific meaning is organised through grammatical resources that allow events and relations to be represented as entities. This enables classification and reasoning across propositions, but it also changes how readers must reconstruct experience from language. For primary education, the challenge is to support access to these resources while preserving the scientific relationships they make available.

An explanation of evaporation illustrates the problem. The expression “evaporation causes cooling” packages two processes into nouns and relates them through a causal verb. A more expanded explanation may describe water particles leaving a surface and explain how this affects the energy of the water that remains. The expanded account is longer and may contain more clauses, yet its sequence can make the causal relationship more visible. Conversely, replacing “evaporation” with “drying” does not necessarily reconstruct the missing mechanism. Neither the shortest text nor the text containing the most familiar words is automatically the most educationally useful. The relevant question is what an explanation enables a particular learner to understand, given the knowledge and linguistic resources that learner brings to the reading situation.

Generative artificial intelligence introduces a new means of producing such explanations. A teacher can request a paragraph for an eight-year-old, a worksheet for a mixed-ability class, or an explanation of a phenomenon observed during an experiment. Related applications include AI-generated fairy tales for primary physics teaching^[2] and ChatGPT-supported production of physics experiment worksheets for primary teachers^[3]. Storytelling provides a relevant context for examining

generated language addressed to children, although its narrative purposes and conventions differ from those of the explanatory prose investigated here. These applications make the language of generated material a practical issue: instructions, explanatory passages, and questions all influence what children attend to and how they describe what happens. The possibility of quickly generating alternative versions is attractive, but the educational value of those versions depends on their scientific and linguistic characteristics rather than their fluency alone.

Recent research provides reasons to examine those characteristics systematically. Berber Sardinha ^[4] identifies differences between human-authored and generated texts through multidimensional analysis. Reinhart *et al.* ^[5] likewise document grammatical and rhetorical patterns that distinguish model outputs from human writing across communicative settings. These findings support the examination of generated language as an empirical object. They do not, however, establish that every generated primary science explanation is excessively dense, nor that all models respond in the same way to audience instructions. Transferring findings from adult genres to explanations for children requires a design that explicitly controls the intended audience, scientific content, text purpose, and conditions of generation.

The present article develops that design while clarifying the theoretical issue it is intended to investigate. Register accommodation is understood as a change in language use that is functionally responsive to a specified communicative situation. Lexical substitution denotes a more limited alteration in word choice that can occur without comparable changes in grammatical organisation or explanatory relationships. The distinction is useful for evaluating an instruction such as “explain this for an eight-year-old,” because that instruction may produce shorter words, a more conversational tone, or direct address without making the underlying scientific process easier to reconstruct. These possible outcomes should be treated as competing empirical possibilities rather than as established properties of ChatGPT.

The article is conceptual and methodological. It does not report a completed corpus, statistical tests, classroom observations, or learning outcomes. Its distinctive contribution is the integration of register analysis with proposition-level scientific evaluation in a research design specifically concerned with primary science explanations. Building on research into register alignment and theoretically informed text adaptation ^[6, 7], the framework brings together clause-based lexical density, contextual analysis of nominalisation, causal explicitness, and reference clarity, while independently assessing scientific accuracy and coverage of essential propositions. This integration makes it possible to distinguish changes in linguistic organisation from changes in scientific meaning without treating either as direct evidence of pupil comprehension. The proposed corpus will examine thirty primary science topics, three audience conditions, repeated generations, and a matched human-authored comparison. A complementary prompt intervention will compare age-based audience specification with additional register guidance. The resulting protocol provides a basis for testing the framework and identifies the learner evidence required before textual differences can be interpreted as educational benefits.

The theoretical synthesis is purposive rather than a

systematic review. It connects foundational accounts of scientific language with selected corpus and educational studies, placing particular emphasis on research published from 2024 to 2026. Bibliographic details and reported findings are checked against publisher records, available abstracts, and source publications. No exhaustive database search, screening flow, or pooled effect estimate is claimed. The literature is used to establish conceptual distinctions and methodological requirements, while the validity of the proposed framework remains an empirical question.

Four research questions organise the initial investigation. The first concerns the lexical, grammatical, and discourse profiles of generated explanations addressed to primary learners. The second concerns whether audience age and the register-guidance package produce coordinated changes across these profiles under a common stipulated prior-knowledge description for each topic. The third concerns how core-generated and register-guided explanations separately compare with human-authored texts selected for corresponding audiences, explanatory purposes, and scientific content. The fourth concerns variation across topics and repeated generations within the documented model version, supplemented by an examination of prompt wording in a prespecified subset. Comparisons across model versions constitute a subsequent replication phase, while effects on pupil understanding require a separate learner study. This division aligns the research questions with the evidence available from each phase.

2. Register and Scientific Meaning in Primary Education

Register links linguistic choices to the circumstances in which language is used. In systemic functional linguistics, field concerns the activity and subject matter, tenor concerns participant relationships, and mode concerns the role of language and the organisation of communication. For the proposed study, field includes the science topic and explanatory purpose, tenor includes the relationship between an explanatory author and an imagined child reader, and mode includes a written explanation intended to stand largely on its own. These dimensions provide a reason to examine coordinated linguistic patterns. Specifying an audience is potentially a change in the communicative situation, rather than merely a request to replace vocabulary. A complementary corpus perspective warns against treating register categories as internally uniform. Biber and Egbert ^[8] emphasise variation within recognised textual varieties and discuss the relationship between situational and linguistic characteristics. This perspective matters when comparing model-generated paragraphs with textbook extracts. A textbook explanation is not linguistically equivalent to a workbook instruction, even when both concern the same phenomenon and age group. Nor does the label “science text” guarantee a single profile. The proposed study therefore treats register as an analytical relation between contextual specifications and linguistic configurations, while preserving differences among explanation, definition, recount, and procedural instruction.

Scientific language often condenses experience through nominalisation and other forms of grammatical metaphor. A process expressed in a clause can be represented as a noun, allowing it to become the subject of another proposition. “The object accelerates” can be reconstrued as “the acceleration of the object.” This shift makes it possible to compare accelerations, describe their conditions, and

connect them to other quantities. Halliday and Martin ^[1] explain how such resources contribute to scientific discourse. Their educational significance lies in their productive role as well as their potential difficulty. Removing every nominalisation would also remove resources that learners eventually need to participate in disciplinary reasoning.

The more appropriate educational objective is therefore supported movement between expanded and condensed representations. An explanation can first identify what an object does, then name the process, and subsequently use the technical noun in a carefully established relationship. The noun “acceleration” becomes more accessible when the text has already explained a change in velocity and clarified what is changing. Halliday and Matthiessen ^[9] provide the grammatical basis for analysing such relations. In primary science, the relevant accommodation may involve making a process explicit before introducing its condensed representation, rather than avoiding the condensed representation altogether.

This distinction also prevents nominalisation from being mistaken for a purely morphological property. A word ending in “-tion” is not automatically a difficult instance of grammatical metaphor, and an everyday noun can package a process in a way that requires substantial reconstruction. “A push” can function as a concrete description of an interaction, but it can also stand for an event whose participants and consequences are left implicit. Scientific terms such as “force” or “energy” demand contextual interpretation that cannot be recovered from suffixes alone. Coding must therefore identify what a nominal expression does within a clause and how it relates to the surrounding explanation.

The scientific role of causal relations is equally central. An explanation should indicate not only that two events occur but how they are related, under which conditions, and with what limits. Xie and Zhang ^[10] examine causal logical grammatical metaphors in academic discourse. Their work points to the importance of considering how causal connections themselves can be grammatically packaged. In a primary explanation, “the effect of heating” may refer to a relationship without stating what changes or why. Explicit clauses can make participants and transitions more visible, but causal connectives do not guarantee scientific adequacy. A sentence containing “because” may still supply a circular or erroneous explanation.

The need to maintain scientific meaning constrains the interpretation of accessibility. A conversational text may be easier to decode but scientifically misleading if it attributes intentions to objects or replaces a mechanism with a familiar analogy. A dense sentence may be appropriate when it summarises a relationship already established in an experiment. Kotsis ^[11] discusses how everyday and specialised language can clarify or obscure physics concepts, emphasising the relationship between linguistic choices, conceptual understanding, and scientific communication. This perspective supports examining whether familiar wording preserves the disciplinary distinctions needed to explain a phenomenon. Snow ^[12] identifies academic language as a challenge in learning from science texts. The present framework extends this concern to generated explanations by asking whether their linguistic organisation supports entry into disciplinary meaning, rather than treating disciplinary language as an obstacle that should

simply be removed.

Primary learners are also heterogeneous. Chronological age is an imperfect proxy for reading proficiency, vocabulary knowledge, prior science learning, language background, and familiarity with the context described. An eight-year-old who has observed condensation may possess experiential resources unavailable to an older child who has only encountered the word. Audience age should therefore be understood as an experimental instruction, not a complete learner model. The initial corpus study will investigate how the system responds to age specifications. Claims about suitability for actual pupils will require separate evidence concerning the readers and the learning tasks for which a text is used.

3. Lexical Density and the Limits of Textual Indicators

Lexical density describes the concentration of lexical material in a text or grammatical unit. Ure ^[13] introduced a formulation based on the proportion of lexical items among running words. Under the operational definition proposed here, lexical items include nouns, lexical verbs, adjectives, and lexical adverbs, while grammatical items include articles, pronouns, conjunctions, auxiliaries, and other function words. The Ure-style measure will be expressed as a percentage: $LD_U = 100L/W$, where L denotes lexical tokens and W denotes all word tokens under the stated tokenisation convention. The measure describes the balance between lexical material and grammatical scaffolding; it is not a count of ideas or a direct estimate of understanding.

A clause-based formulation associated with Halliday uses the ranking clause as the denominator. The proposed implementation will calculate $LD_H = L/C$, where C denotes ranking clauses and L denotes lexical tokens in the text, including lexical tokens inside embedded structures. Embedded clauses functioning within a nominal group will not add separate ranking clauses to the denominator. This convention requires manual validation, because a dependency parser does not automatically distinguish ranking clauses from embedded clauses in the systemic functional sense. The result will be reported as lexical items per ranking clause rather than as a percentage, making the difference between the two formulations explicit.

The formulations can respond differently to a revision. Adding conjunctions, pronouns, or explanatory phrases can reduce the lexical percentage without reducing the amount of lexical material per ranking clause. Expanding a nominalised process into a ranking clause can increase the denominator of the clause-based measure while altering the balance of lexical and grammatical items. A passage can consequently become less dense under one definition and remain similar under another. This divergence is potentially informative, but it does not identify a pedagogical mechanism without examining the revised clauses and their scientific content. The protocol will therefore inspect discordant cases rather than collapsing both measures into a single density score.

To and Mahboob ^[14] examine lexical density and nominalisation in reading passages across levels of an English textbook series. Their analysis connects linguistic complexity with textbook progression while also considering science and non-science fields. This study provides a methodological precedent for examining density alongside grammatical resources. Its educational setting differs from the proposed primary science corpus, however,

and its findings should not be treated as age-specific density standards. A locally observed increase across textbook levels does not establish a universal threshold separating suitable from unsuitable science explanations.

Lexical diversity is a separate property. Density concerns the proportion or concentration of lexical tokens, whereas diversity concerns the range of distinct vocabulary used. A text can be lexically dense while repeating a small set of technical terms. Such repetition may be helpful when it maintains stable reference, even though it lowers a diversity index. Conversely, replacing a technical term with several near-synonyms can increase diversity while making the concept harder to track. Nasser and Thompson [15] examine density and diversity in dissertation abstracts, demonstrating the relevance of keeping the two constructs analytically distinct. Their adult academic corpus does not directly establish their relation to children's comprehension.

Raw type-token ratio is sensitive to length and will not be the primary diversity index. The proposed analysis will use moving-average type-token ratio with a predefined fifty-token window [16] and will report the handling of texts shorter than that window. A conventional standardised type-token ratio will be retained as a supplementary measure for sufficiently long passages. Sensitivity analyses will use additional window lengths and an alternative index such as MTLD, drawing on the validation work of McCarthy and Jarvis [17]. These alternatives will help establish whether an apparent audience difference is robust or mainly an artefact of text length and measurement choice.

Conventional readability formulas provide another descriptive perspective. Flesch Reading Ease and Flesch-Kincaid Grade Level are functions of average word or syllable length and sentence length. They can be useful for characterising surface change, but they do not directly identify nominalisation, scientific correctness, referential ambiguity, or omitted causal steps. Crossley *et al.* [18] show the value of richer linguistic features for predicting human readability judgments and reading speed. Importantly, this does not make every conventional formula uninformative; it limits the claims that can be made from any one score. A lower predicted grade level should be reported as a lower formula score, not as demonstrated understanding by younger children.

The effect of nominalisation on readability scores is not predetermined. Condensing clauses can shorten a sentence, but the resulting nouns may be longer, and the text may contain complex modifiers. Different formula components can therefore move in opposite directions. The framework avoids the claim that nominalised prose will necessarily be judged easier by a surface formula. Instead, it proposes comparing the directions and magnitudes of observed changes, then analysing the passages that show disagreement between formula-based readability and register indicators. Such disagreement generates a research question rather than resolving it in advance.

Additional indicators will include nominalisation frequency, clause structure, technical vocabulary coverage, reference clarity, and causal explicitness. Mean T-unit length will measure words per independent clause together with its dependent clauses, while clauses per T-unit will describe clausal elaboration. Neither measure will substitute for the ranking-clause analysis. Technical vocabulary coverage will be defined as the proportion of word tokens belonging to a topic-specific, independently constructed technical lexicon.

Frequency of direct address and rhetorical questions will provide limited evidence about audience positioning. These measures are complementary because they describe different aspects of an explanation, and Table 1 specifies how each will be interpreted.

Table 1: Proposed Linguistic Indicators and Their Interpretation

Indicator	Operationalisation	Interpretive boundary
Lexical density by words	$100 \times \text{lexical tokens} / \text{word tokens}$	Balance of lexical material and grammatical scaffolding; not a comprehension score.
Lexical density by clauses	Lexical tokens / ranking clauses	Content concentration under an explicit clause-coding convention.
Nominalisation	Occurrences per 100 words; supplementary rate per clause	Contextually coded process and quality nominalisations; explained terms distinguished.
Lexical diversity	MATTR with a 50-token window; sensitivity checks	Vocabulary range; term repetition may support stable reference.
Clause organisation	Words per T-unit; clauses per T-unit	Clausal elaboration, interpreted alongside nominal packaging.
Technical vocabulary	Technical tokens / word tokens; explanation status	Disciplinary terminology and its support; fewer terms is not always preferable.
Causal and reference clarity	Independent coding of required relations and referents	Recoverability of explanatory links; reviewed separately from scientific accuracy.
Surface readability	Flesch Reading Ease; Flesch-Kincaid Grade Level	Formula-based surface estimates; no claim of validated age suitability.
Audience positioning	Direct address and rhetorical questions per 100 words	Interpersonal cues; friendliness alone does not establish accommodation.

Note. All measures describe text properties. Scientific adequacy and pupil comprehension require separate evaluation. MATTR = moving-average type-token ratio.

No measure in Table 1 is proposed as an automatic quality score. The framework rejects a compensatory interpretation in which a very low readability grade could offset a scientific error or unclear reference. It also rejects an undifferentiated “simplicity” composite whose weighting obscures differences between vocabulary, grammar, and content. A profile of indicators is preferable because it preserves the possibility that adaptation occurs unevenly. This profile will guide qualitative inspection and hypothesis testing while keeping scientific adequacy as a separately evaluated condition.

4. Evidence on Generated Language and Audience Adaptation

Research on generated language has moved beyond whether readers can recognise an artificial source. Corpus studies increasingly ask whether outputs reproduce the functional patterns associated with particular communicative settings. Goulart *et al.* [19] compare situational and linguistic characteristics of undergraduate writing and AI-generated assignments, while Mizumoto *et al.* [20] examine linguistic fingerprints in comparisons involving EFL writing. These studies support the use of systematic linguistic comparison, but they do not provide a direct basis for diagnosing the readability of explanations for primary pupils. The present

article uses them as methodological and contextual evidence, with their population and genre boundaries retained.

Multidimensional studies provide a broader reason to investigate register. Berber Sardinha ^[4] compares generated and human-authored texts across co-occurring linguistic features. Reinhart *et al.* ^[5] examine variation in grammatical and rhetorical styles across several model systems and text contexts. Their results indicate that generated prose may have recognisable stylistic preferences and may not reproduce human variation consistently. The relevant implication for primary science is a hypothesis: an audience label may coexist with a persistent informational style. Establishing whether this occurs in school science requires comparing multiple topics and prompts rather than extrapolating from one illustrative output.

Demir and Egbert ^[6] provide more closely related evidence by examining register alignment in generated academic texts across journal articles and textbooks in biology and history. Their analysis reports a mixed pattern, with less successful alignment in textbook contexts than in journal article writing. This is especially relevant because educational texts must serve pedagogical purposes, even when their subject matter overlaps with research prose. Nevertheless, the study's textbook comparisons should not be treated as direct evidence concerning eight-year-old readers. It strengthens the rationale for controlling both genre and intended audience in a primary science investigation.

Studies of argumentation offer a further caution about collapsing distinct linguistic resources. Jiang and Hyland ^[21] compare lexical bundles in argumentative essays, and Jiang and Hyland ^[22] examine metadiscursive nouns. These resources contribute to organising propositions and positioning an argument, but their frequency is not equivalent to the frequency of scientific nominalisations. A noun that labels an argument as a "claim" has a different discourse function from a noun that reconstrues a physical process as "evaporation." The proposed codebook will preserve this distinction, preventing evidence about academic metadiscourse from being used as a direct measure of explanatory compression.

Audience adaptation is also more responsive to prompting than a simple failure account would suggest. Huynh and McNamara ^[7] investigate adaptation of science texts to hypothetical reader profiles across several models. Their work identifies inconsistent alignment in an initial experiment and improved alignment when prompts are augmented with theoretically informed guidance. This finding supports a comparison between minimal audience specification and more explicit adaptation instructions. It also indicates that educational evaluation should examine the actual prompt-model combination. A conclusion about an age-only prompt cannot automatically be extended to a carefully supported adaptation workflow.

Watanabe ^[23] examines attempts to ease text readability using different model and prompt conditions. Successful change is reported for the grade-targeted conditions in the second experiment, whereas a more general request in the first experiment did not yield comparable evidence. The measures include linguistic sophistication, cohesion, and holistic readability. It would therefore be inaccurate to characterise this research as relying only on traditional word- and sentence-length formulas. The present framework complements such work by adding explicit systemic

functional coding and scientific content review, rather than claiming that prior research has ignored grammatical or discourse features altogether.

Newer research clarifies why prompt specification and temporal stability should be included in the design. Crosthwaite *et al.* ^[24] show that metadiscourse-inducing prompts can alter distributions and rhetorical realisations in generated argumentative writing. Huynh and McNamara ^[25] investigate repeated personalised science texts and model updates, reporting short-term consistency in their tested conditions alongside differences across versions. Together, these studies caution against both assuming unrestricted instability and assuming stable model behaviour. Repeated sampling within a version and comparisons across documented versions address different questions and should be analysed separately.

The construction of comparable corpora remains essential. Milička *et al.* ^[26] present English and Czech corpora generated to support comparison with established human corpora. Their resource demonstrates the value of preserving genre coverage and comparability when analysing generated language. The proposed study adopts the general principle of explicit documentation and matched comparisons, while using a smaller corpus tailored to primary science. It does not assume that observations made in English will generalise to Greek or other languages, because tokenisation, morphology, clause structure, technical vocabulary, and schooling conventions all influence the meaning of the indicators.

Science education scholarship also provides a basis for treating language as more than a delivery channel. Tang ^[27] synthesises language and literacy research to inform investigations of generative AI in science education. Kalantzis and Cope ^[28] discuss literacy in the context of artificial intelligence, broadening attention to the production and use of texts. Kotsis ^[29] similarly frames AI integration in science education as a pedagogical and epistemological issue. These perspectives motivate evaluation of what generated explanations ask learners to do with scientific meaning, rather than evaluation limited to how polished those explanations appear.

5. A Framework for Register Accommodation

The proposed framework distinguishes four relations that are often combined under the term "simplification." Surface adjustment concerns word length, sentence length, and familiar vocabulary. Grammatical accommodation concerns the distribution of processes across clauses and nominal groups. Explanatory accommodation concerns the visibility of participants, causal steps, conditions, and referential links. Disciplinary progression concerns how a text introduces and subsequently uses the language needed for scientific reasoning. These relations can co-occur, but one does not logically entail another. A model can produce a friendly tone while preserving grammatical compression, or expand processes while introducing a scientifically inaccurate mechanism.

Register accommodation will be operationalised as a coordinated, context-sensitive change in linguistic organisation, supported by close analysis of meaning. The operationalisation does not require every indicator to decrease. Introducing an essential technical noun may increase measured lexical density while improving conceptual precision. Repeating the noun after an accessible

explanation may reduce lexical diversity while strengthening reference. Similarly, adding a causal clause may increase sentence length yet make a relationship explicit. The interpretation must therefore consider whether the observed changes serve the stipulated audience and explanatory purpose, rather than defining accommodation as universally lower scores.

Lexical substitution will describe cases in which the main evidence concerns changes in word choice or lexical sophistication while grammatical packaging and explanatory relationships remain substantially similar. This is a descriptive category, not an automatic negative judgment. Replacing an unfamiliar nonessential word may be a valuable revision when the explanation is otherwise appropriate. The concern arises when such a revision is presented as comprehensive audience adaptation. An explanation may then be labelled suitable for a child even though the same hidden process relations remain. The proposed analysis will distinguish limited lexical change from broader accommodation without presuming that limited change is useless.

A further category is scientific content alteration. An output may omit a necessary condition, add an irrelevant

mechanism, or transform an analogy into an assertion about how nature works. Such changes should not be interpreted merely as shifts along a linguistic continuum. They alter the explanation's scientific commitments. The framework therefore places content preservation and accuracy alongside linguistic analysis as necessary checks. It permits a generated explanation to be classified as linguistically accommodated but scientifically inadequate, or scientifically adequate but linguistically demanding. The two judgments answer different questions.

Table 2 provides constructed examples of these possibilities. The passages are written solely to illustrate the framework; they are not ChatGPT outputs, corpus observations, or evidence about the performance of any model. They address the familiar observation that water on a plate gradually disappears into the surrounding air. The condensed version names processes while leaving relationships compressed. The vocabulary-only version uses more familiar expressions but retains abstract packaging. The expanded version identifies what happens and connects it to the technical term. The misleading version attributes an intention to water, showing why familiar language cannot replace scientific review.

Table 2: Constructed Illustrations of Adaptation and Scientific Adequacy

Formulation	Illustrative passage	Analytical interpretation
Condensed formulation	Evaporation produces a reduction in the amount of water on the plate.	Processes are packaged as nouns; the transformation is not unpacked.
Vocabulary substitution	Drying leads to a drop in the amount of water on the plate.	More familiar wording retains abstract packaging and an underspecified process.
Expanded formulation	Some water at the surface changes from a liquid into a gas called water vapour. The water vapour mixes with the surrounding air, so less liquid water remains on the plate. This change is called evaporation.	Processes, referents, and the technical label are made explicit; suitability still requires learner evidence.
Misleading familiar formulation	The water wants to leave the plate, so it escapes into the air.	Accessible wording assigns intention to water; scientific review identifies the problem.

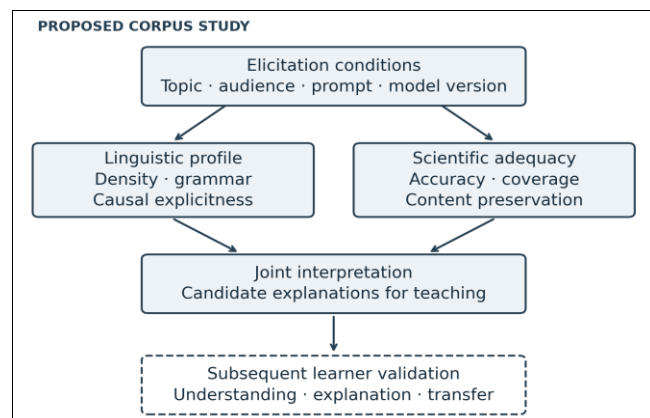
Note. These passages were constructed for illustration. They are not model outputs or empirical observations. No density values or comprehension effects are inferred from them.

The expanded passage should not be interpreted as a fully validated explanation for every primary learner. It presupposes some understanding of air and introduces the idea of a gas that cannot usually be seen. A classroom teacher may need to establish these ideas through observations and representations before the paragraph becomes useful. The example demonstrates a linguistic possibility rather than an age-specific learning effect. This boundary is central to the framework: a text can make a relationship more explicit without establishing that a particular child has understood it.

The framework also treats causal explicitness as relational rather than purely lexical. Counting “because,” “therefore,” or “so” provides a useful starting point, but close coding will examine whether the connective links a stated cause to a scientifically relevant consequence. A circular statement such as “the water evaporates because evaporation occurs” is linguistically explicit but explanatorily empty. Likewise, an explanation of friction that says objects “want to stop” may be memorable but replaces interaction with intention. Kotsis ^[30] documents anthropocentric reasoning in pre-

service teachers’ explanations of weight, friction, buoyancy, and air resistance, demonstrating the relevance of examining such reasoning in physics explanations. Although that study does not establish whether ChatGPT produces comparable explanations, it supports including anthropocentric attribution as a category for scientific scrutiny. Scientific review must therefore assess the propositions linked by the grammar, not simply whether a causal marker is present.

Fig 1 represents the relationships among elicitation conditions, linguistic profiles, scientific adequacy, and learner validation. The first two analytical branches are complementary: one describes how meaning is grammatically organised, while the other evaluates whether the explanation preserves scientifically appropriate relationships. Their convergence supports selection of candidate teaching texts, but it does not demonstrate learning. A subsequent investigation involving pupils must establish whether those texts support interpretation, explanation, or transfer. The figure deliberately separates the proposed corpus study from that later validation phase.



Note. Solid boxes describe the proposed text-based investigation. Topic, audience, and prompt condition vary within the initial design; model version is documented and held constant. Variation across model versions belongs to a subsequent replication phase. The dashed box represents a separate study involving learners. Arrows express analytical relationships, not observed causal effects.

Fig 1: Relations between elicitation, text evaluation, and learner validation

Three testable expectations follow from the framework. Age-only instructions may produce stronger changes in surface indicators than in grammatical packaging. Explicit register guidance may produce additional shifts in clause structure and causal explicitness, although it may also affect length or content coverage. Finally, the magnitude of these effects may differ across topics, because explaining a process, defining a quantity, and describing an interaction place different demands on language. These expectations are provisional. Findings showing coordinated accommodation under age-only prompts, no additional benefit from register guidance, or strong stability across topics would also be informative and must remain possible outcomes.

6. The Proposed Corpus and Generation Protocol

The primary corpus will consist of explanations for thirty topics mapped to an identified primary science curriculum. A provisional allocation will include ten topics in physical science, ten in life science, and ten in Earth and environmental science. The sampling frame will document curriculum references, expected prior knowledge, intended explanatory purpose, and essential scientific propositions for each topic. The final topic list will be fixed before the main generation phase. A small pilot using topics outside the main corpus will refine instructions and annotation rules, ensuring that methodological decisions do not depend on inspecting the eventual study outcomes.

The audience conditions will specify readers aged approximately eight, ten, and twelve years. These ages represent experimentally controlled labels, not claims about uniform developmental stages or international grade equivalence. The prompts will describe an imagined English-reading pupil and will not include identifiable pupil information. The study will focus initially on written explanatory prose rather than dialogue, multiple-choice tasks, or procedural worksheets. This restriction makes the mode and communicative purpose more comparable across conditions. Subsequent studies may broaden the design, but they should do so explicitly rather than mixing fundamentally different tasks within an apparently homogeneous corpus.

The proposed core design will generate five independent outputs for each topic and audience condition using one documented model version and one fixed prompt template. This yields $30 \times 3 \times 5 = 450$ generated explanations. The number is a design target, not a completed sample or a power guarantee. Each request will use an independent context, and the order of topic–audience combinations will be randomised or counterbalanced. Keeping generation contexts independent will prevent earlier explanations from influencing subsequent ones through accumulated conversation history. Repetition will address variation among outputs generated under the same documented conditions.

A matched human-authored corpus will provisionally contain two texts per topic and audience condition, giving 180 comparison texts if suitable material can be identified. Sources will include publisher-produced primary science materials and other educational publications with identifiable intended audiences and authorship histories. Selection will prioritise explanatory purpose, topic coverage, language, and approximate length. Published educational material will not be assumed to be scientifically adequate solely because it is human-authored. Kazantzidou and Kotsis^[31] identified misrepresentations of the nature, location, and role of ozone in children’s trade books, providing a precedent for evaluating written representations against the scientific consensus. The human-authored passages will therefore undergo the same scientific accuracy and content-coverage evaluation as the generated explanations. Texts whose AI provenance cannot be reasonably assessed will be flagged and, where uncertainty is substantial, excluded from the principal human-authored comparison. “Human-authored” will be treated as a documented provenance claim rather than inferred from linguistic appearance.

Human-text selection will follow a documented eligibility form completed before linguistic outcomes are examined. Each passage must have an identifiable educational source, documented intended audience, and the same explanatory purpose as the corresponding generated-text condition. Audience matching will use publisher age recommendations or a documented mapping between the source’s curriculum year and the target age; wording alone will not establish audience suitability. Eligible passages will normally contain 150–250 words, providing a stated tolerance around the generated-text target of 180–220 words. Shorter or longer passages will be reserved for supplementary analysis. Minimum conceptual overlap will require the same focal phenomenon, principal entities, and explanatory relationship; a definition of evaporation will therefore not qualify as a match for an explanation of evaporative cooling. Coverage of each essential proposition will be recorded separately, allowing incomplete coverage to remain visible. Passages that require an accompanying illustration or omitted surrounding text to recover the central explanation will be excluded from the principal comparison. Among eligible candidates, selection will prioritise audience correspondence and proximity to the target length, using a prespecified tie-breaking procedure. Cells without suitable material will remain unmatched, and comparison analyses will report their coverage and include a sensitivity analysis restricted to fully matched topic–audience cells.

The core prompt will use the following fixed template: “Write a continuous science explanation in English for a

pupil aged [age] about [topic]. The pupil has the following prior knowledge: [brief curriculum-aligned description]. Explain [specified phenomenon or relationship] and include these essential scientific points: [content specification]. Aim for 180–220 words. Use prose without headings, lists, equations, or dialogue.” For each topic, the prior-knowledge description, explanatory purpose, and essential scientific propositions will remain identical across the three age conditions and both prompt conditions. The prior-knowledge description will specify a minimal background plausibly shared across the intended age range, with its suitability reviewed during the pilot. Consequently, the audience contrast will estimate the effect of changing the age label under a common stipulated knowledge profile, rather than differences between complete age-specific learner profiles. Outputs outside the requested word range will be retained and flagged unless a prespecified technical-failure rule applies. No output will be regenerated solely because its length or wording is inconvenient.

The register-guided condition will reproduce the core prompt verbatim and append the following instruction: “Describe key processes in explicit clauses before introducing terms that name those processes. Make clear which objects, organisms, or substances each statement concerns, and use pronouns only when their referents are clear. Explain the relevant causes and consequences rather than merely naming their relationship. Define necessary scientific terms when first introduced and use them consistently thereafter. Do not assign intentions or desires to physical objects or non-intentional processes. Preserve the specified scientific points and the requested word range.” This condition will contain five repetitions per topic and age, adding 450 explanations. The principal contrast will estimate the effect of this combined guidance package; it will not identify the separate contribution of any individual instruction. All completed prompt instances will be archived with their corresponding outputs. Table 3 summarises the proposed corpus components.

Table 3: Proposed corpus composition

Component	Design	Target texts	Conditions
Core generated corpus	30 topics × 3 ages × 5 repetitions	450	Fixed core prompt; one documented model version.
Register-guided generated corpus	30 topics × 3 ages × 5 repetitions	450	Core prompt plus explicit grammatical and explanatory guidance.
Human-authored comparison	30 topics × 3 ages × 2 texts	180	Target subject to comparable texts and documented provenance.
Prompt wording sensitivity	Prespecified stratified subset	To be fixed	Paraphrased core prompts; supplementary analysis.
Model version extension	Replication using documented later versions	Future phase	Distinct from repetition within one version.

Note. Numbers are proposed collection targets, not observed samples or power estimates. Ages are approximately 8, 10, and 12 years. The principal planned corpus totals 1,080 texts before supplementary arms.

Prompt wording sensitivity will be examined in a prespecified subset rather than ignored. Two semantically comparable paraphrases of the core instruction will be developed before collection and applied to a stratified selection of topics. Their wording will avoid introducing new scientific content or explicit register guidance. This supplementary arm will test whether conclusions are overly dependent on one phrasing. It will be analysed separately from the principal core-versus-guided contrast. The exact subset size and repetition count will be finalised using resource constraints and simulation, then recorded before the supplementary outputs are generated.

The study will document model identifier, accessible version or snapshot information, platform, generation dates, system-level instructions under researcher control, temperature, top-p, maximum output length, and seed where supported. Unavailable or unexposed settings will be recorded as unavailable rather than inferred. The protocol does not assume that any particular parameter is exposed by a conversational interface. A versioned API is preferable when it permits more complete documentation, but interface-based collection remains possible if its reproducibility limitations are reported. Automatic model selection or routing, if applicable, would require disclosure because it weakens the interpretation of a single-model corpus.

All requests and raw outputs will be preserved before cleaning. Refusals, incomplete answers, unexpected lists, excessive length, and apparent factual errors will remain part of the generation record. The main linguistic corpus will apply clearly stated eligibility rules, while a separate compliance analysis will report these departures. No output will be silently improved before evaluation. For human texts, source, publication date, intended audience, permission status, and selection rationale will be recorded. These records allow the corpus to be understood as a documented sample of educational explanations rather than a collection of favourable examples.

7. Annotation and Scientific Evaluation

Automated preprocessing will produce tokenisation, sentence segmentation, lemmatisation, part-of-speech tags, and dependency annotations using a documented software environment. The choice of tool and language model will be recorded with version identifiers. Preprocessing will preserve a link between each token and the original text so that coding decisions can be checked. Hyphenated expressions, contractions, numerals, and symbols will follow a written tokenisation policy. Because technical vocabulary and child-directed writing can challenge automated taggers, the pipeline will be treated as an aid to annotation rather than an unquestioned source of ground truth.

Ranking clauses will be identified through a combination of automated candidate detection and manual coding. The codebook will specify the treatment of coordinated clauses, relative clauses, nonfinite clauses, ellipsis, and embedded clauses functioning within nominal groups. For example, the relationship between “water changes into vapour” and “the change of water into vapour” must be interpreted

consistently when calculating clause-based density. Ambiguous cases will be flagged rather than resolved differently in different audience groups. The final report will describe these conventions and include representative annotated examples so that readers can understand what was counted.

Nominalisation coding will distinguish process nominalisations, quality nominalisations, and relational or metadiscursive nouns. A separate contextual judgment will identify whether a nominalisation compresses scientific content that the passage subsequently unpacks. This prevents a count from treating a technical term introduced after explanation as equivalent to an unexplained abstraction. The primary nominalisation measure will be occurrences per one hundred word tokens, supplemented by occurrences per ranking clause. A morphology-based candidate search may improve efficiency, but contextual validation will be required, and its recall will be checked against manually annotated passages.

Technical vocabulary will be defined independently of the model outputs. For each topic, reviewers will construct a lexicon from curriculum documents, scientific reference materials, and appropriately selected educational sources. Multiword terms will be represented explicitly, while polysemous terms such as “work,” “force,” “current,” and “power” will require contextual coding. Terms will also be classified according to whether they are explained, exemplified, or left unexplained. This distinction matters because identical coverage percentages can conceal very different opportunities for comprehension. The analysis will not assume that reducing technical vocabulary is desirable when doing so removes the language needed to identify the scientific concept.

Causal evaluation will distinguish expression, recoverability, and scientific correctness for each required relation specified in the topic’s content framework. Expression will be coded using nominal categories: stated within a clause, expressed across clauses, implied, or absent. Recoverability will be coded separately as fully recoverable, partially recoverable, or not recoverable. A fully recoverable relation will identify the relevant participants and the asserted cause–consequence connection without requiring the reader to supply an unstated explanatory step. Scientific correctness will be assessed independently against the topic specification, allowing a clearly expressed but incorrect connection to receive high recoverability and an adverse accuracy judgment. Absence will be recorded as missing coverage rather than automatically classified as an incorrect assertion. The primary causal outcome will be the proportion of required relations coded as fully recoverable; the complete category distributions will remain available for secondary analysis. Reference clarity will concern whether pronouns and noun phrases can be connected to their intended participants. The pilot codebook will include correct but unclear, clear but incorrect, incomplete, and ambiguous examples, and its definitions will be fixed before main-corpus annotation.

At least two trained coders will independently annotate a stratified subset containing both sources, all audience conditions, both principal prompt conditions, and all science domains. The initial target will be twenty per cent of eligible texts, subject to feasibility and adequate coverage of infrequent categories. Coders will first train on pilot texts that are outside this reliability subset. They will then code

independently before discussing disagreements. Agreement will be calculated on the unreconciled annotations, and reconciled codes will be used only after the reliability analysis has been completed. This sequence avoids overstating agreement by calculating it after consensus.

Reliability statistics will match the measurement level. Krippendorff’s alpha will be considered for categorical annotations, with an explicitly stated distance function for ordinal categories and a missing-data policy. Hayes and Krippendorff^[32] provide a methodological basis for this approach. Continuous counts or rates will use an intraclass correlation specification appropriate to the coder design, with confidence intervals and the distinction between consistency and absolute agreement made explicit. Koo and Li^[33] discuss reporting choices for intraclass correlations. Token- or span-level categories will additionally report appropriate precision, recall, and overlap measures where a simple text-level coefficient would conceal segmentation disagreements.

Scientific accuracy will be evaluated separately by reviewers with relevant science education expertise. This distinction is supported by Kazantzidou and Kotsis^[34], whose analysis of children’s books available in Greece identified explanations of ozone layer depletion that did not consistently conform to the scientific consensus. Their findings provide a precedent for examining the accuracy of causal relationships in child-directed texts independently of their intended educational purpose. Reviewers will compare each passage with the topic’s essential propositions and assess incorrect assertions, misleading analogies, omitted conditions, and unsupported attribution of agency. They will also record whether simplification changes the explanatory goal. For instance, a passage may correctly say that a plant needs light while failing to explain the role of light in the intended account of photosynthesis. Such a passage is not necessarily false, but it may be incomplete for the specified purpose. Distinguishing error from incomplete coverage is necessary for interpreting adaptation.

Reviewers will receive anonymised texts in random order, with source labels and prompt condition removed where feasible. The intended audience and topic must remain available because adequacy is contextual, but avoidable cues to provenance will be minimised. Blinding cannot be guaranteed: distinctive wording or publication conventions may suggest source. The study will report this limitation and may ask reviewers to record provenance guesses separately. Disagreements about scientific judgments will be preserved for agreement analysis before adjudication. The final dataset will retain both initial judgments and adjudicated decisions so that uncertainty remains visible.

The evaluation will also separate raw outputs from teacher-revised versions. The first study concerns what the model produces under specified conditions. If a teacher edits a passage after generation, that creates a different object: a human–AI production workflow. Such revisions may be educationally valuable, but they should be logged and analysed in a subsequent arm rather than substituted for raw outputs. This distinction preserves the interpretability of model comparisons and creates a clear route for later research on whether professional review can improve scientific and linguistic adequacy together.

8. The Proposed Analytical Strategy

The explanation will be the primary observational unit.

Tokens, clauses, and repeated outputs are nested within explanations and topic–audience conditions, so treating them as independent cases would inflate apparent precision. Multiple human texts from the same publication or author may also share features unrelated to the experimental audience specification. The statistical analysis will therefore use multilevel models where the data support them. Bates *et al.* [35] describe the implementation of linear mixed-effects modelling in lme4, providing a practical reference for analysing clustered observations without reducing the corpus to unstructured group means.

The principal generated-text analysis will estimate the effects of audience condition, prompt condition, and their interaction, with topic included as a grouping factor. Topic-specific audience effects will be represented through random slopes when estimable and justified by the design. Science domain will be considered as a fixed factor, because the three selected domains are substantively defined rather than a large random sample of domains. Repetition within a topic–audience–prompt cell will contribute to estimates of residual generation variability. The final random-effects structure will be checked for singularity and supported by design-based reasoning rather than selected solely to obtain favourable significance levels.

Outcome distributions will determine model families. Nominalisation counts can be analysed with an appropriate count model and a word-count offset; overdispersion will be assessed before choosing a Poisson specification. Proportions such as technical coverage may require binomial-type models or other suitable treatment when dependence among tokens prevents a simple independent-binomial interpretation. Continuous density measures will be checked for skew and residual patterns, with transformations or robust alternatives considered where necessary. Reporting means alone would conceal potentially important variation, so distributions and confidence intervals will accompany model estimates.

Text length requires particular care because it can be both a constraint and a consequence of adaptation. The primary analysis will describe actual lengths and estimate overall differences under the fixed requested range. A secondary analysis will adjust for actual length or use matched-length subsets to examine whether particular indicators remain different. The adjusted estimate answers a conditional question and must not replace the total effect without explanation. Including length as a covariate can remove part of the very adaptation process being studied. Both perspectives will therefore be labelled clearly, and departures from the requested length range will be reported.

The human–generated comparison will include a three-level production-condition factor comprising human-authored, core-generated, and register-guided-generated explanations, together with audience and their interaction. Planned contrasts will compare each generated condition separately with the human-authored reference corpus. Topic will be represented as a shared grouping factor, with human publication or author included where the sampling structure supports estimation. Generated repetitions and human passages will not be treated as matched one-to-one observations merely because they occupy the same topic–audience cell. Analyses will preserve the distinction between experimentally elicited outputs and selected published passages, and will examine residual differences in length and proposition coverage through clearly labelled

sensitivity analyses. Production-condition estimates will describe differences between the sampled corpora; comparisons with human texts will not establish causal effects of authorship or universal differences in educational quality.

The analysis of accommodation will combine prespecified contrasts with a structured interpretation rubric. Limited lexical substitution will describe changes towards more familiar wording without corresponding changes in process expression, reference clarity, or the recoverability of required causal relations. Such judgments will be supported by contextual examination of vocabulary choices rather than inferred from lexical diversity or technical vocabulary coverage alone. Broader register accommodation will require convergent evidence of change in at least two of three dimensions: grammatical organisation, reference clarity, and causal explicitness. Grammatical evidence will include processes expressed through explicit clauses or nominalisations whose meanings are unpacked; referential evidence will concern clearer identification of participants; and causal evidence will concern more recoverable relationships between scientifically relevant propositions. These judgments will be recorded separately from scientific accuracy and content coverage, allowing a passage to show linguistic accommodation while remaining scientifically inadequate. Increased density or sentence length will not disqualify a passage when textual evidence demonstrates clearer explanatory organisation. Cases with conflicting evidence will be classified as mixed or indeterminate. The rubric will be refined using pilot texts and fixed before main-corpus annotation; interpretations will report effect sizes, uncertainty, and representative textual evidence.

Claims about stability will distinguish variance from absence of a trend. Repeated outputs can show no systematic difference by repetition order while still varying substantially within a prompt cell. The study will estimate within-cell variation and compare it with variation attributable to audience and prompt condition. Topic-level variance and, in a later extension, version-level differences will be reported separately. Intraclass correlations or variance ratios may help summarise clustering, but they must be interpreted alongside the underlying outcome scales. A statistically detectable audience effect may have limited practical value if individual outputs vary widely around the expected profile.

Three primary outcomes will be specified: lexical items per ranking clause, process and quality nominalisation occurrences per one hundred words, and the proportion of required causal relations coded as fully recoverable. Metadiscursive nouns will be excluded from the primary nominalisation count. For generated explanations, each primary model will include audience, prompt condition, their interaction, and science domain as fixed effects, with a topic random intercept and design-supported random slopes where estimable. Clause-based density will initially use a linear mixed-effects model, subject to distributional checks. Nominalisation counts will use a Poisson or negative-binomial mixed model with the logarithm of word count as an offset, according to the prespecified overdispersion assessment. Fully recoverable relations will be analysed as a grouped binomial response comprising recoverable and remaining required relations, with additional dispersion accommodated where necessary. Exact contrasts, diagnostic procedures, and fallback specifications will be registered

after the pilot and before main collection. Holm adjustment will apply across the three primary tests of the average guided-versus-core prompt contrast; audience and interaction tests will form separately specified confirmatory families. Other linguistic indicators will support secondary analyses. Pilot reliability problems will prompt clarification and retraining before the codebook is fixed. If main-corpus reliability remains inadequate under the preregistered criterion, the affected coded outcome will be reported as exploratory rather than replaced retrospectively by a more favourable measure.

Power and precision will be evaluated through simulation after a pilot supplies plausible variance ranges. Thirty topics and five repetitions are starting points that balance topical coverage and feasibility, not a claim that the design is adequately powered for every interaction. Simulations will vary topic heterogeneity, residual variation, and expected prompt effects to identify contrasts that can be estimated with useful precision. Where resources are limited, increasing topical breadth may be more informative than repeatedly sampling a few topics. The final design will state which effects it is intended to estimate and which require a larger follow-up study.

Sensitivity analyses will examine alternative tokenisation choices, diversity windows, treatment of ambiguous nominalisations, topic imbalance, and exclusion of scientifically inadequate passages. The full corpus will remain the primary description of generated behaviour, because removing inaccurate outputs can make a model appear more suitable than it is. Analyses restricted to scientifically adequate explanations will answer the additional question of linguistic variation conditional on adequacy. Missing comparisons, parser failures, and annotation uncertainty will be documented. Results will be illustrated with distribution plots, estimated contrasts, and carefully chosen excerpts after collection; no simulated results are presented in this article.

9. Educational Implications

The framework has immediate implications for how teachers review generated science material. A request for a specified age produces an intended audience label, but it does not certify a text's adequacy for pupils in a particular classroom. Teachers can examine whether the passage identifies participants, describes observable changes, explains relevant relationships, and introduces scientific terminology through recoverable meaning. This approach makes evaluation more specific than asking whether the wording "sounds simple." It also supports revision: a difficult sentence can be unpacked into processes, then reconnected to the technical formulation learners need to encounter.

Technical terms should be introduced strategically rather than treated as avoidable complications. "Evaporation," "friction," and "condensation" allow children to connect different instances of a phenomenon and participate in later science learning. Removing these terms entirely may create an accessible story without providing a stable conceptual label. A teacher may therefore use a progression from observation to explicit explanation to condensed terminology. The proposed framework makes this progression visible and distinguishes purposeful technical vocabulary from unnecessary lexical elaboration. A reduction in technical coverage is desirable only when it

removes dispensable burden while retaining essential scientific distinctions.

The framework is particularly relevant to experimentation. A generated worksheet can ask pupils to carry out steps successfully while leaving the explanatory meaning of the observations implicit. Kotsis^[3] identifies a context in which ChatGPT can support worksheet production, but linguistic evaluation remains necessary for the explanatory passages embedded in such resources. A teacher reviewing a paragraph after an experiment should ask whether it connects the observed change to the intended concept and whether its referents match the objects children actually manipulated. This links textual accommodation to the material context of learning rather than treating the paragraph as an isolated linguistic product.

Teacher education must also address the evaluation of scientific content. Stylos *et al.*^[36] investigate scientific literacy among Greek pre-service primary teachers. Their work provides a relevant basis for discussing the knowledge teachers bring to evaluating explanations, although it does not measure their ability to analyse AI-generated prose. Kotsis^[37] proposes a human-centred framework for professional development that emphasises teachers' critical evaluation and adaptation of AI resources in science teaching. This orientation supports incorporating scientific and linguistic scrutiny into teacher preparation. Professional preparation could combine scientific review with sentence-level comparison: future teachers can inspect alternative explanations, identify omitted conditions, and discuss how grammatical choices make a mechanism visible or obscure. They could subsequently revise generated passages while documenting how their changes preserve scientific meaning and clarify explanatory relationships. Such activities would require empirical evaluation before claims are made about their effects on teaching competence.

The role of the teacher extends beyond approving text. A teacher can connect written explanations with oral discussion, diagrams, observations, and questions that reveal pupils' interpretations. Tang^[27] supports attention to the language and literacy dimensions of science education when investigating generative AI. Within this perspective, a generated paragraph becomes one component of a meaning-making activity. Its linguistic profile may influence that activity, but classroom interaction can unpack terms, repair reference, and expose misunderstandings that an automated text score cannot detect. Consequently, evaluating the text and evaluating its use are related but distinct research tasks. Inclusive practice also requires care in how audience profiles are specified. A single age label cannot represent differences in language experience, reading proficiency, or prior learning. Teachers may need versions that preserve the same scientific propositions while varying the amount of explanation, reference support, and contextualisation. Kasneci *et al.*^[38] discuss educational opportunities and challenges associated with large language models, including the importance of responsible use. The present proposal gives one focused application of that concern: differentiated wording should preserve a shared disciplinary goal, while uncertainty about learner suitability remains open to classroom evidence.

There is a risk of narrowing access to disciplinary language if generated "easy" texts are assigned permanently to pupils considered weaker readers. An accessible version should

provide a route towards the concepts and formulations used in the wider curriculum. Otherwise, differentiation can become exclusion from the language of science. This is a conceptual implication of the framework rather than an observed effect of current practice. It suggests examining whether adapted explanations establish bridges between everyday and scientific expression, and whether pupils are offered opportunities to use the technical language after its meaning has been introduced.

Finally, automated evaluation should support professional judgment without replacing it. A dashboard showing lexical density or readability can help locate passages for review, but it cannot decide whether the relevant mechanism is correct or whether an analogy is appropriate for the classroom. Kotsis ^[29] frames AI integration in science education as involving pedagogical and epistemological choices. Here, those choices concern which linguistic demands to reduce, which scientific distinctions to retain, and how to connect explanatory prose with inquiry. The value of a linguistic profile is that it makes some of these choices inspectable rather than presenting a single opaque suitability label.

10. Limitations and Directions for Learner Validation

The principal limitation is that the proposed framework has not yet been tested using the specified corpus. Its categories and measures are theoretically motivated, but their empirical coherence and usefulness remain to be established. A future study may find that clause-based density contributes little beyond other indicators, that some coding distinctions are unreliable, or that different science topics require different profiles. These outcomes would warrant revising the framework. The present article therefore offers operational commitments and falsifiable expectations, not a validated instrument for certifying educational text.

A second limitation concerns the gap between linguistic description and comprehension. A lower density value does not demonstrate that a child has constructed a better explanatory model. A longer passage may provide useful causal support but exceed the learner's attention or working-memory resources. The influence of connectives on young readers, examined by Cain and Nash ^[39], illustrates why local linguistic resources must be considered in relation to actual processing. The planned corpus can identify candidate differences, but it cannot establish how these differences interact with prior knowledge, reading skill, task demands, or teacher support.

The human-authored comparison also has limitations. Published educational texts vary in quality, editorial practices, intended contexts, and dependence on accompanying diagrams. Some may be written to be read aloud, whereas others presume independent reading. Human authorship is not a guarantee of scientific or linguistic adequacy. Matching and contextual documentation can reduce these problems, but cannot eliminate them. The comparison should therefore function as a situated reference distribution and a source of contrasting examples, not an unquestionable gold standard against which every generated passage is judged.

English-only analysis restricts transfer. Greek and other languages differ in morphology, lexical formation, word order, and educational terminology, all of which affect computational indicators and grammatical analysis. A translated prompt may also elicit a different style rather than

a language-equivalent version of the same output. The multilingual corpus work of Milička *et al.* ^[26] supports the importance of constructing comparable resources, but does not remove these analytical challenges. A Greek-language extension would require a validated tokenizer, a language-appropriate technical lexicon, revised clause coding where needed, and independently selected human comparison texts.

Model version and prompting define another boundary. The study will describe a documented system under a specified set of instructions and collection dates. It will not establish enduring properties of the ChatGPT brand or of generative AI as a whole. Huynh and McNamara ^[25] show why comparisons across updates are necessary when evaluating temporal consistency, while Crosthwaite *et al.* ^[24] demonstrate the methodological relevance of prompt specificity. Replication should therefore preserve exact prompts and distinguish reanalysis of an archived corpus from new generation using a later system. Both are useful, but they answer different questions.

A subsequent learner study should test a small set of scientifically adequate explanations that differ in the targeted linguistic properties. Selection should maintain essential propositions and explanatory purpose while documenting unavoidable differences in length and vocabulary. Pupils' prior knowledge and reading proficiency would need assessment before exposure. Random assignment or a counterbalanced within-pupil design could help identify text-condition effects, with appropriate modelling of pupil, classroom, topic, and text dependencies. The goal would be to test whether the proposed linguistic accommodations affect interpretation, rather than whether pupils prefer one style of paragraph.

Outcome measures should extend beyond immediate factual recall. Learners could explain the phenomenon in their own words, identify the participants and causal relationship, evaluate an incorrect account, and apply the explanation to a related situation. Oral explanation may reveal understanding that a written response obscures, while delayed tasks could examine retention. A qualitative component could trace where pupils struggle with nominal groups, pronoun reference, or causal links. Such evidence would permit the framework to be revised around demonstrated learning processes, including cases in which a theoretically promising adaptation produces no measurable benefit.

Any investigation involving pupils would require an appropriate ethics review, consent and assent procedures, and safeguards suitable to its setting. The corpus study proposed here involves no pupil participation or personal learner data. Permissions for human-authored materials and decisions about corpus redistribution would still need to be established before collection. The present article claims no completed ethical approval, registered protocol, or available dataset. Its implementation should record the applicable institutional determination and material permissions rather than assuming that a text-only design automatically settles every governance question.

11. Transparency and Research Integrity

Transparency is especially important when the research object is generated language. The study should archive the elicitation protocol, raw outputs, preprocessing code, software versions, codebook, and adjudication records, subject to lawful redistribution of comparison materials.

Where human texts cannot be shared, the repository should provide source identifiers, selection rules, derived features, and permitted excerpts sufficient to support scrutiny. Generated texts must retain their provenance and collection metadata. A table of aggregate scores without the associated prompts and texts would be inadequate for checking whether the study has measured the claimed linguistic distinctions.

The analysis plan should be registered before collection of the main corpus. Registration will distinguish confirmatory questions from exploratory examination and state eligibility rules, primary outcomes, model specifications, and planned sensitivity analyses. It should not be treated as a prohibition on learning from unexpected patterns; justified departures can be reported transparently. The protocol in this article is a proposal rather than an already registered commitment. Implementation will need to resolve remaining choices, particularly the final topic list, human comparison availability, and precision targets, without representing those future decisions as completed work.

Reporting will preserve the distinction between illustrative material and collected evidence. Table 2 contains constructed examples, Table 3 specifies collection targets, and the framework's expectations remain hypotheses to be tested. Subsequent publications will identify which protocol components were implemented and document any departures from the registered plan.

Use of generative AI in manuscript development should be disclosed in accordance with the target journal's policy. In preparing this draft, generative AI assisted with the development of the conceptual framework, prose drafting, literature checking, and document preparation. No empirical observations or statistical results were generated for presentation as research findings. The author retains responsibility for the arguments, citations, illustrative examples, and final wording of the submitted manuscript. This disclosure reflects the scope of assistance in the present draft rather than describing it solely as language correction.

12. Conclusion

The language of generated science explanations should be evaluated in relation to the scientific meaning it organises and the learners it is intended to support. Familiar vocabulary, short sentences, and a friendly tone can contribute to accessibility, but they do not establish that causal relationships are explicit, technical terms are recoverable, or the explanation is scientifically adequate. Register provides a framework for examining these relationships because it connects linguistic patterns with communicative purpose, audience, and mode. Lexical density contributes to this framework as a descriptive measure whose interpretation depends on complementary grammatical and content analysis.

This article has developed a conceptual distinction between limited lexical substitution and broader register accommodation, while avoiding the assumption that every desirable adaptation must reduce every linguistic indicator. It has specified a proposed corpus design, an explicit comparison of age-only and register-guided prompts, a plan for repeated sampling and matched human texts, and procedures for annotation reliability and scientific review. These provisions make the research questions testable. They also define the scope of the evidence that the study can produce: findings about generated text under documented

conditions, rather than direct proof of learning or enduring claims about all models.

The proposed programme ultimately aims to connect linguistic analysis with science education decisions. Its immediate contribution is to make teacher review and future corpus research more precise. Its longer-term contribution will depend on studies involving learners that test whether particular ways of expressing processes, connecting causes, and introducing terminology support understanding and transfer. Until that evidence is available, a generated explanation should be treated as a candidate instructional resource whose scientific and linguistic properties require examination. The appropriate goal is accessible entry into disciplinary meaning, with the language of science introduced through explanation rather than removed from it.

13. References

- Halliday MAK, Martin JR. Writing science: Literacy and discursive power. Pittsburgh: University of Pittsburgh Press, 1993.
- Kotsis KT. Artificial intelligence creates fairy tales for physics teaching in primary education. *Eur J Open Educ Learn Stud.* 2024; 9(1):1-16. Doi: 10.46827/ejoe.v9i1.5250
- Kotsis KT. ChatGPT develops physics experiment worksheets for primary education teachers. *Eur J Educ Stud.* 2024; 11(5):1-20. Doi: 10.46827/ejes.v11i5.5274
- Berber Sardinha T. AI-generated vs human-authored texts: A multidimensional comparison. *Appl Corpus Linguist.* 2024; 4(1):100083. Doi: 10.1016/j.acorp.2023.100083
- Reinhart A, Markey B, Laudénbach M, Pantusen K, Yurko R, Weinberg G, *et al.* Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proc Natl Acad Sci USA.* 2025; 122(8):e2422455122. Doi: 10.1073/pnas.2422455122
- Demir NY, Egbert J. Register alignment of ChatGPT-generated academic texts. *Appl Corpus Linguist.* 2026; 6(1):100174. Doi: 10.1016/j.acorp.2025.100174
- Huynh L, McNamara DS. GenAI-powered text personalization: Natural language processing validation of adaptation capabilities. *Appl Sci.* 2025; 15(12):6791. Doi: 10.3390/app15126791
- Biber D, Egbert J. What is a register? Accounting for linguistic and situational variation within and outside of-textual varieties. *Register Stud.* 2023; 5(1):1-22. Doi: 10.1075/rs.00004.bib
- Halliday MAK, Matthiessen CMIM. Halliday's introduction to functional grammar. 4th ed. London: Routledge, 2014. Doi: 10.4324/9780203431269
- Xie Y, Zhang Y. Enhancing academic discourse: Features and functions of causal logical grammatical metaphors. *Linguist Lit Stud.* 2024; 12(3):29-45. Doi: 10.13189/lls.2024.120302
- Kotsis KT. The language of physics: Shaping conceptual understanding and scientific communication. *Bull Lang Lit Stud.* 2025; 2(2):18-28. Doi: 10.59652/bls.v2i2.581
- Snow CE. Academic language and the challenge of reading for learning about science. *Science.* 2010; 328(5977):450-452. Doi: 10.1126/science.1182597
- Ure J. Lexical density and register differentiation. In: Perren GE, Trim JLM, editors. *Applications of linguistics.* Cambridge: Cambridge University Press,

- 1971, 443-452.
14. To V, Mahboob A. Complexity of English textbook language: A systemic functional analysis. *Linguist Hum Sci.* 2019; 13(3):264-293. Doi: 10.1558/lhs.31905
 15. Nasser M, Thompson P. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assess Writ.* 2021; 47:100511. Doi: 10.1016/j.asw.2020.100511
 16. Covington MA, McFall JD. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *J Quant Linguist.* 2010; 17(2):94-100. Doi: 10.1080/09296171003643098
 17. McCarthy PM, Jarvis S. MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav Res Methods.* 2010; 42(2):381-392. Doi: 10.3758/BRM.42.2.381
 18. Crossley SA, Skalicky S, Dascalu M. Moving beyond classic readability formulas: New methods and new models. *J Res Read.* 2019; 42(3-4):541-561. Doi: 10.1111/1467-9817.12283
 19. Goulart L, Matte ML, Mendoza A, Alvarado L, Veloso I. AI or student writing? Analyzing the situational and linguistic characteristics of undergraduate student writing and AI-generated assignments. *J Second Lang Writ.* 2024; 66:101160. Doi: 10.1016/j.jslw.2024.101160
 20. Mizumoto A, Yasuda S, Tamura Y. Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Appl Corpus Linguist.* 2024; 4(3):100106. Doi: 10.1016/j.acorp.2024.100106
 21. Jiang F, Hyland K. Does ChatGPT argue like students? Bundles in argumentative essays. *Appl Linguist.* 2025; 46(3):375-391. Doi: 10.1093/applin/amae052
 22. Jiang F, Hyland K. Metadiscursive nouns in academic argument: ChatGPT vs student practices. *J Engl Acad Purp.* 2025; 75:101514. Doi: 10.1016/j.jeap.2025.101514
 23. Watanabe M. ChatGPT, can you make this text easier?: Investigating the utility of LLMs in easing text readability. *Int J Artif Intell Educ.* 2025; 35(6):3922-3942. Doi: 10.1007/s40593-025-00520-7
 24. Crosthwaite P, Sun SA, Jiang FK. Prompt sensitivity in AI writing: A methodological caution for corpus-based analysis of AI-generated language. *Mod Lang J.* 2026. Epub ahead of print. Doi: 10.1111/modl.70089
 25. Huynh L, McNamara DS. Evaluation of linguistic consistency of LLM-generated text personalization using natural language processing. *Electronics.* 2026; 15(6):1262. Doi: 10.3390/electronics15061262
 26. Milička J, Marklová A, Cvrček V. AI Brown and AI Koditex: LLM-generated corpora comparable to traditional corpora of English and Czech texts [preprint]. *arXiv.* 2025. arXiv:2509.22996. Doi: 10.48550/arXiv.2509.22996
 27. Tang KS. Informing research on generative artificial intelligence from a language and literacy perspective: A meta-synthesis of studies in science education. *Sci Educ.* 2024; 108(5):1329-1355. Doi: 10.1002/sce.21875
 28. Kalantzis M, Cope B. Literacy in the time of artificial intelligence. *Read Res Q.* 2025; 60(1):e591. Doi: 10.1002/rrq.591
 29. Kotsis KT. Artificial intelligence as a catalyst for changes in university-level science education. *EIKI J Eff Teach Methods.* 2025; 3(3):99-109. Doi: 10.59652/jetm.v3i3.618
 30. Kotsis KT. Anthropocentric reasoning of weight, friction, buoyancy, and air resistance among pre-service teachers. *EIKI J Eff Teach Methods.* 2025; 3(4). Doi: 10.59652/jetm.v3i4.668
 31. Kazantzidou D, Kotsis KT. Representation of the ozone layer in children's trade books about ozone layer depletion: An analysis of written texts in Greece. *Interdiscip J Environ Sci Educ.* 2023; 19(1):e2302. Doi: 10.29333/ijese/12847
 32. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas.* 2007; 1(1):77-89. Doi: 10.1080/19312450709336664
 33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016; 15(2):155-163. Doi: 10.1016/j.jcm.2016.02.012
 34. Kazantzidou D, Kotsis KT. Ozone layer depletion in children's books available in Greece: Examining accuracy in the representation of causes of ozone layer depletion in texts. *Child Lit Educ.* 2024; 55(4):567-585. Doi: 10.1007/s10583-023-09524-0
 35. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015; 67(1):1-48. Doi: 10.18637/jss.v067.i01
 36. Stylos G, Siarka O, Kotsis KT. Assessing Greek pre-service primary teachers' scientific literacy. *Eur J Sci Math Educ.* 2023; 11(2):271-282. Doi: 10.30935/scimath/12637
 37. Kotsis KT. A human-centred framework for teacher professional development in AI-supported science teaching. *J Artif Intell Technol Dev.* 2026; 2(3):9-30. Doi: 10.59324/jaitd.2026.2(3).02
 38. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, *et al.* ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ.* 2023; 103:102274. Doi: 10.1016/j.lindif.2023.102274
 39. Cain K, Nash HM. The influence of connectives on young readers' processing and comprehension of text. *J Educ Psychol.* 2011; 103(2):429-441. Doi: 10.1037/a0022824