Artificial Intelligence Chemistry 4 (2026) 100118

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

ChemPro: A progressive chemistry benchmark for Large Language Models

Aaditya Baranwal, Shruti Vyas *

University of Central Florida, 4000 Central Florida Blvd. Orlando, FL, 32816, United States of America

ARTICLE INFO

Keywords:
Chemistry benchmark
Large Language Models (LLMs)
Progressive difficulty
Curriculum-aligned evaluation
Scientific reasoning
Articulation complexity
Multiple choice questions (MCQs)
Numerical questions
Performance degradation
Subfield analysis

ABSTRACT

We introduce ***ChemPro***, a progressive benchmark with 4100 natural language question-answer pairs in Chemistry, across 4 coherent sections of *difficulty* designed to assess the proficiency of Large Language Models (LLMs) in a broad spectrum of general chemistry topics. We include Multiple Choice Questions and Numerical Questions spread across fine-grained information recall, long-horizon reasoning, multi-concept questions, problem-solving with nuanced articulation, and straightforward questions in a balanced ratio, effectively covering *Bio-Chemistry*, *Inorganic-Chemistry*, *Organic-Chemistry* and *Physical-Chemistry*. ***ChemPro*** is carefully designed analogous to a student's academic evaluation for basic to high-school chemistry. A gradual increase in the question *difficulty* rigorously tests the ability of LLMs to progress from solving basic problems to solving more sophisticated challenges.

We evaluate 45+7 state-of-the-art LLMs, spanning both open-source and proprietary variants, and our analysis reveals that while LLMs perform well on basic chemistry questions, their accuracy declines with different types and levels of complexity. These findings highlight the critical limitations of LLMs in general scientific reasoning and understanding and point towards understudied dimensions of difficulty, emphasizing the need for more robust methodologies to improve LLMs.

1. Introduction

LLMs have demonstrated strong capabilities in language, coding, mathematics, and physics [1–4]. However, robust scientific reasoning remains under-evaluated at the foundational level, where conceptual understanding, numerical reasoning, and multi-step problem solving are required.

Existing LLMs fail to adequately meet the requirements of significantly assisting professional scientific research. The LLMs are not high-agency in the context of emergent properties and abilities for science. Often called the *central science* [19], chemistry requires linguistic comprehension alongside symbolic manipulation (e.g., equations), multi-step calculations, and reasoning over reaction mechanisms, and is not bound under a single formal language [20,21]. Despite its importance, chemistry has received comparatively less benchmark attention [21,22], leaving a gap in evaluating foundational proficiency beyond rule-bound domains like programming and mathematics [23–25].

To bridge this gap, we introduce ***ChemPro*** (Fig. 2), a progressive benchmark designed to evaluate LLM chemistry proficiency via a curriculum-aligned progression of questions [26]. ChemPro[1] consists of 4100 questions sourced from standardized materials including competitive exams [27], textbooks [28], and quizlets [29], spanning Inorganic, Organic, Physical and Bio-chemistry.

We conduct a comprehensive evaluation of state-of-the-art LLMs [30–33] on ChemPro, analyzing their performance across a diverse set of chemistry topics (Fig. 1). Our experiments span 45 different models, exploring proprietary and open-source variants with varying model sizes.

Coverage and provenance. ChemPro uses educational provenance (NCERT and JEE Mains) to enforce *difficulty ordering* while spanning high-school chemistry; results are reported per tier and subfield.

2. Related work

General-Purpose and STEM Benchmarks: Foundation models have led to specialized benchmarks across domains. In mathematics, MATH [34] evaluates symbolic reasoning, while programming benchmarks like HumanEval [35] assess code generation. Multi-domain benchmarks

* Correspondence to: Eng-1, 408A, University of Central Florida, 12760 Pegasus Dr, Orlando, FL 32816, USA.
E-mail address: shruti@ucf.edu (S. Vyas).

[1] We use *CP* as acronym for ChemPro everywhere in the paper.

https://doi.org/10.1016/j.aichem.2026.100118
Received 14 November 2025; Received in revised form 13 March 2026; Accepted 9 April 2026
Available online 13 April 2026

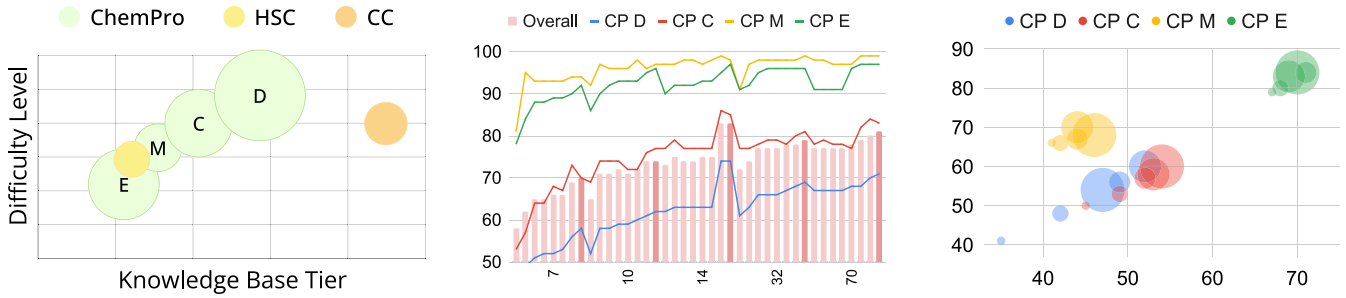

Fig. 1. ***Overview of ChemPro. Left:*** A comparison with existing benchmarks. *Y-axis* represents LLM difficulty (lower mean MCQ accuracy across models = harder). *X-axis* represents academic succession (Elementary to Graduate/Expert), rendered as a continuum because real-world curricula overlap across grade boundaries. *Bubble size* represents question count. **E**, **M**, **C**, and **D** are ChemPro's **E**asy, **M**edium, **C**hallenging, and **D**ifficult sections (axis derivation details in supplementary). ***Center:*** Performance (Accuracy, y-axis) of all 40 open-source models evaluated on ChemPro MCQs showing the impact of model-size (x-axis) on performance (lines are Performance on individual ChemPro sections and columns is the overall average performance). ***Right:*** Exact-match accuracy (x-axis) vs Tolerance-based accuracy (y-axis) on ChemPro Numerical for all 40 open-source models (bubble size represents model parameter count).

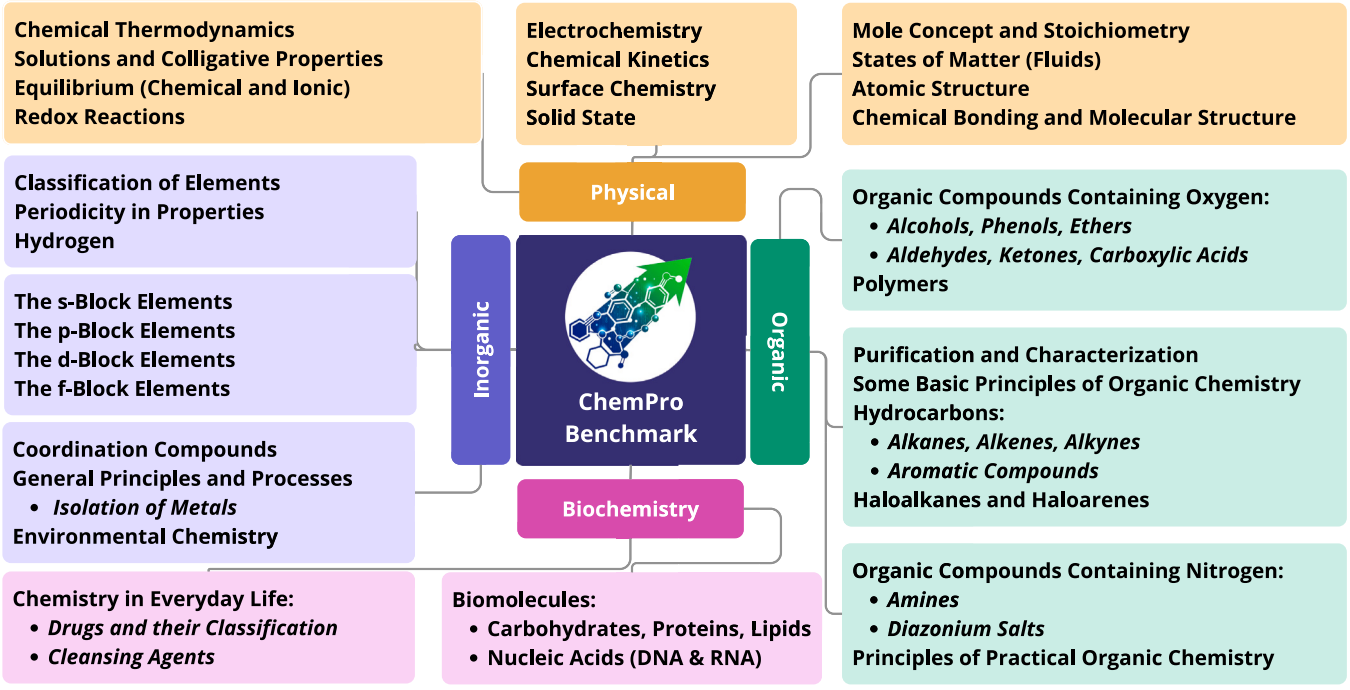

Fig. 2. ***ChemPro benchmark structure:*** The benchmark spans four chemistry subfields (Biochemistry, Inorganic, Organic, Physical Chemistry) across four sections of difficulty (CP_E, CP_M, CP_C, CP_D) with balanced distribution of MCQs and numerical problems. Complete category distribution details are provided in Appendix.

include MMLU [7] and GPQA [8] targeting graduate-level questions. For scientific reasoning, ARC [5] provides elementary to high-school science questions but lacks chemistry-specific depth. SciBench [10] evaluates college-level scientific problem-solving but focuses on undergraduate-to-graduate content. JEEBench [9] evaluates high-school problems but lacks systematic difficulty progression within subjects.

Chemistry-Specific Benchmarks: Recent chemistry-focused models like ChemLactica [36] and Llama-Chem [37] demonstrate domain expertise but lack systematic evaluation across curriculum-aligned difficulty levels. Molecular-focused benchmarks include SMolInstruct [11], MolInstruct [12], and CACTUS [13], which target expert-level capabilities rather than fundamentals. Research Literature focused benchmarks include ScholarChemQA [14] and ChemBench [15], are focused on advanced topics and lack structure for generalizability and systematic difficulty assessment, works like MolVision [38] try to address molecular chemistry from vision modality.

Critical Limitations: Current chemistry evaluation suffers from three fundamental gaps: (1) **Difficulty annotation subjectivity**-most benchmarks use broad categorizations lacking verifiable educational grounding; (2) **Inadequate foundational coverage**-existing benchmarks target specialized tasks, neglecting systematic elementary-to-high-school evaluation; (3) **Limited assessment modalities**-few benchmarks combine conceptual understanding (MCQs) with computational reasoning (numerical problems). Structured comparison between relevant (STEM and chemistry) benchmarks and ChemPro is provided in Table 1.

ChemPro's Positioning: ChemPro addresses these limitations through: **Source-aligned difficulty provenance** tied to established curricula (NCERT) and examinations (JEE), providing verifiable educational ordering; **Comprehensive E-HS focus** systematically evaluating foundational chemistry within educational boundaries; **Dual-mode assessment** with multiple choice questions and numerical problems for comprehensive evaluation of LLMs (see Table 2).

Table 1
Comparison with chemistry and related science benchmarks. ChemPro uniquely provides *source-aligned difficulty provenance* and comprehensive *elementary-to-high-school* coverage with dual assessment modes. *Knowledge Tier*: Ex = Expert/Research, UG = Undergraduate, HS = High School, E = Elementary.

Benchmark	Topics	Format	Knowledge Tier	Progressive
ARC [5]	Science	MCQs	E-HS	No
ScienceQA [6]	Science	MCQs	E-HS	No
MMLU [7]	Multi-domain	MCQs	UG-Ex	No
GPQA [8]	Multi-domain	MCQs	Graduate-Ex	No
JEEBench [9]	STEM	MCQ+Open	HS-Advanced	No
SciBench [10]	STEM	MCQ+Open	UG-Advanced	No
SMolInstruct [11]	Chemistry	Instruction	UG-Research	No
MolInstruct [12]	Chemistry	Instruction	UG-Research	No
CACTUS [13]	Chemistry	Agent Tasks	UG-Research	No
ScholarChemQA [14]	Chemistry	Literature	Advanced-Research	No
ChemBench [15]	Chemistry	MCQ+Open	UG-Graduate	No
ChemLLMBench [16]	Chemistry	Open+Code	Advanced-Research	No
RESTEEM [17]	Chemistry	Educational	E-UG	No
HS Chemistry [18]	Chemistry	MCQs	HS	No
College Chemistry [18]	Chemistry	MCQs	UG-Graduate	No
ChemPro (Ours)	**Chemistry**	**MCQ + Numerical**	**E-HS**	**Yes**

Table 2
Performance across ChemPro MCQ difficulty levels: Top 3 performing models across ChemPro MCQ sections by model size category. Performance metrics show accuracy scores for each section with progressive difficulty from Easy to Difficult. Note the systematic performance degradation (highlighted in red) as difficulty increases. (*P: Proprietary*).

Best Models	Size	$\mathcal{CP}_E$	$\mathcal{CP}_M$	$\mathcal{CP}_C \downarrow$	$\mathcal{CP}_D \Downarrow$
HomerCreativeAnvita-Mix-Qw7B [39]		0.89	0.93	0.67	0.53
Falcon3-Jessi-v0.4-7B-Slerp [40]	7B	0.90	0.94	0.73	0.56
Falcon3 7B Instruct [41]		0.92	0.94	0.70	0.58
MT-Merge4-gemma-2-9B [42]		0.93	0.98	0.72	0.60
falcon3-10b-tensopolis-v1 [43]	10B	0.95	0.96	0.76	0.61
Falcon3 10B Instruct [44]		0.96	0.97	0.77	0.62
Lamarckvergence-14B [45]		0.93	0.98	0.77	0.63
Phi-4-Model-Stock-v4 [46]	14B	0.95	0.99	0.80	0.67
Luminis-PHI-4 [47]		0.97	0.98	0.85	0.74
PathFinderAi3.0 [48]		0.96	0.98	0.78	0.67
Qwen2.5-32B-Instruct-abliterated-v2 [49]	32B	0.96	0.98	0.80	0.68
Rombos-LLM-V2.5-Qwen-32B [50]		0.96	0.99	0.81	0.69
shuttle-3 [51]		0.97	0.99	0.82	0.68
Homer-v1.0-Qwen2.5-72B [52]	70B	0.97	0.99	0.84	0.70
Rombos-LLM-V2.5-Qwen-72B [50]		0.97	0.99	0.83	0.71
OpenAI o1-mini [53]		0.97	0.98	0.83	0.75
OpenAI o3-mini [54]	P	0.96	0.99	0.84	0.75
OpenAI o1 [53]		0.97	0.99	0.85	0.76

3. ChemPro benchmark

ChemPro is a curriculum-aligned progressive benchmark designed to systematically evaluate LLM chemistry proficiency across elementary-to-high-school difficulty levels.

3.1. Benchmark definition and curation framework

Formulation of ChemPro. Let $\mathcal{L} : \mathcal{Q} \to \mathcal{A}$ represent an LLM function mapping questions to answers. We define ChemPro through a systematic curation process:

We define source spaces $\mathcal{S} = \{\mathcal{S}_E, \mathcal{S}_M, \mathcal{S}_C, \mathcal{S}_D\}$ where:

$$\mathcal{S}_E = \{s \in \text{Web} : \text{difficulty}(s) = \text{Elementary}\} \quad (1)$$

$$\mathcal{S}_M = \{s \in \text{NCERT}_{9-10} : \text{grade}(s) \in [9, 10]\} \quad (2)$$

$$\mathcal{S}_C = \{s \in \text{NCERT}_{11-12} : \text{grade}(s) \in [11, 12]\} \quad (3)$$

$$\mathcal{S}_D = \{s \in \text{JEE}_{2020-2024} : \text{year}(s) \in [2020, 2024]\} \quad (4)$$

For each question $q \in \mathcal{Q}$, validation $\mathcal{V} : \mathcal{Q} \to \{0, 1\}$:

$$\mathcal{V}(q) = \mathcal{V}_{\text{source}}(q) \wedge \mathcal{V}_{\text{expert}}(q) \wedge \mathcal{V}_{\text{AI}}(q) \quad (5)$$

where $\mathcal{V}_{\text{source}}$, $\mathcal{V}_{\text{expert}}$, and $\mathcal{V}_{\text{AI}}$ represent source verification, expert review, and AI-assisted validation respectively.

Verification stages. $\mathcal{V}_{\text{source}}$ enforces provenance/traceability, $\mathcal{V}_{\text{expert}}$ enforces correctness and unambiguity after textual adaptation, and $\mathcal{V}_{\text{AI}}$ performs automated consistency checks (formatting, deduplication, leakage flags); full criteria are in the supplementary.

The final benchmark construction follows:

$$\mathcal{CP} = \bigcup_{i \in \{E,M,C,D\}} \mathcal{CP}_i \text{ where } \mathcal{CP}_i = \{q \in \mathcal{S}_i : \mathcal{V}(q) = 1\} \quad (6)$$

Assessment Structure: Each $\mathcal{CP}_i$ is partitioned into:

$$\mathcal{CP}_i = \mathcal{M}_i \cup \mathcal{N}_i \text{ with } \mathcal{M}_i \cap \mathcal{N}_i = \emptyset \quad (7)$$

where $\mathcal{M}_i$ denotes multiple-choice questions and $\mathcal{N}_i$ denotes numerical problems, ensuring coverage of both conceptual understanding and computational reasoning.

The alignment with educational sources ensures:

$$\mathcal{CP}_E \prec \mathcal{CP}_M \prec \mathcal{CP}_C \prec \mathcal{CP}_D \quad (8)$$

where $\prec$ denotes curriculum-verified difficulty progression.

Source-Aligned Difficulty Provenance. Unlike existing benchmarks with loosely defined difficulty levels, our sections are intrinsically tied to established educational sources that represent statistical

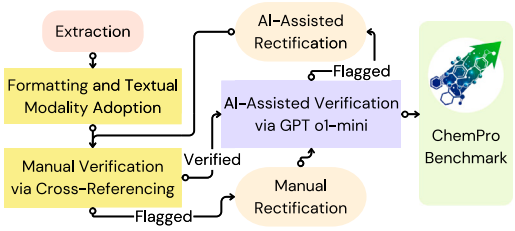

Fig. 3. ***ChemPro benchmark curation process.*** Visual workflow showing the systematic approach for creating ChemPro benchmark, from source collection across different difficulty tiers to quality validation and final dataset compilation. The process ensures source-aligned difficulty provenance while maintaining rigorous quality standards through multiple validation layers (both AI and Human).

consensus across thousands of educators and years of curriculum refinement: $\mathcal{CP}_E$: Web sources, online quizlets, questionnaires, covering elementary concepts. $\mathcal{CP}_M$: intermediate understanding. $\mathcal{CP}_C$: advanced high-school level. $\mathcal{CP}_D$: competitive-level problem solving.

Curriculum Consistency Validation. JEE Mains examination follows the official NCERT syllabus as mandated by the National Testing Agency, ensuring that $\mathcal{CP}_C$ and $\mathcal{CP}_D$ questions assess identical conceptual boundaries. The systematic performance differences between these sections (average 13-point accuracy drop from $\mathcal{CP}_C$ to $\mathcal{CP}_D$ across all models) therefore reflect variations in question formulation complexity rather than conceptual scope expansion.

Articulation complexity. We use this term to denote formulation-induced complexity (e.g., chained reasoning steps, conversions, cross-condition integration, and multi-concept coupling) beyond the concept label; tiering acts as a provenance-based proxy (details in supplementary).

Statistical Reliability Over Annotator Judgments. This source-aligned approach provides complexity measures that are orders of magnitude more reliable than individual annotator ratings, eliminating subjectivity inherent in expert and now-popular AI annotations while ensuring difficulty progression reflects genuine educational complexity.

Performance Validation. The systematic 13-point accuracy drop between $\mathcal{CP}_C$ and $\mathcal{CP}_D$ across all 45 evaluated models, despite curriculum equivalence, provides empirical evidence that observed difficulty stems from formulation complexity rather than conceptual scope differences.

3.2. Deduplication and uniqueness validation

We conducted rigorous quality assurance: (1) Cross-benchmark deduplication with existing datasets using n-gram similarity analysis confirmed minimal overlap of only 6 questions (0.15% of our dataset); (2) Advanced leakage detection using GPT-4o [55] with four distinct methodologies: *prefix completion testing* (systematic truncation at various points to detect completion patterns), *paraphrasing detection* (generation of semantic equivalents), *content modification analysis* (numerical and formula alterations), and *reverse engineering* (concept-based generation); more details in supplementary. This comprehensive analysis identified a mere $\approx$ 8% potential exposure; (3) $\mathcal{CP}_D$ questions from JEE (2020–2024) are inherently resistant to memorization due to multi-step reasoning requirements.

3.3. Dataset composition

ChemPro contains 4100 questions: $\mathcal{CP}_D$ (2315, 56.4%), $\mathcal{CP}_C$ (665, 16.2%), $\mathcal{CP}_M$ (335, 8.2%), and $\mathcal{CP}_E$ (795, 19.3%). Each question underwent three-fold verification including source verification, expert review, and AI-assisted validation (detailed procedures in Fig. 3).

Difficulty distribution. ChemPro intentionally contains more $\mathcal{CP}_D$ items to stress-test competitive, multi-step formulations; we therefore report results per-tier and per-subfield.

Textual Adaptation. Chemistry visuals were systematically converted to text using: standardized LaTeX equations, IUPAC nomenclature for structures, numbered sequences for reaction mechanisms, and structured descriptions for diagrams. These use representations commonly found in educational materials [17], minimizing linguistic complexity while preserving visual provenance.

3.4. Evaluation framework and experimental setup

For MCQs: $\text{Acc}_{\mathcal{M}} = \frac{1}{|\mathcal{M}|}\sum_{i=1}^{|\mathcal{M}|}\mathbb{I}[f(\mathcal{L}, q_i) = a_i]$
For numerical problems:

- **Exact Match**: $\text{Acc}_{exact} = \frac{1}{|\mathcal{N}|}\sum_{i=1}^{|\mathcal{N}|}\mathbb{I}[\hat{y}_i = y_i]$
- **Tolerance**: $\text{Acc}_{tol} = \frac{1}{|\mathcal{N}|}\sum_{i=1}^{|\mathcal{N}|}\mathbb{I}[|\hat{y}_i - y_i| \leq \theta \cdot \frac{|y_i|+|\hat{y}_i|}{2}]$

where $\theta = 0.1$ is the tolerance threshold. This approach distinguishes between conceptual understanding (tolerance-based) and computational precision (exact match).

Answer Verification Numericals are scored on the parsed numeric value extracted from `FINAL ANSWER` (not raw string equality). Questions specify integer/two-decimal rounding and units; if units are omitted, the expected SI-unit answer is in an appropriate range (details in supplementary).

Model Selection. We include 45+7 LLMs: (1) 40 open-source models from the OpenLLM Leaderboard representing top performers across five parameter scales (7B, 10B, 14, 32B, 70B) to capture scaling effects in general-purpose language models, 5 chemistry corpora trained language models [56,57] and 2 latest general opensource models [58]; (2) State-of-the-art proprietary models: GPT-3.5-Turbo, GPT-4o, o1-mini, o3-mini, and o1 representing current commercial capabilities; (3) ChemCrow agentic framework for analysis vertical scaling with tool augmentation. Focus on general-purpose models addresses a critical gap: while specialized chemistry models proliferate for narrow research tasks, foundational chemistry remains underserved. Complete model list in Appendix.

Evaluation Protocol. All models evaluated pass@1. We probe COT, Self-Critique and ICL finalizing an empirically robust system and wraparound prompts (separate for MCQs and Numericals) which are consistent for all models (system prompt is appended to user-input with wraparound prompt for reasoning models). ChemCrow is evaluated as-is.

Availability and licensing

The ChemPro benchmark is constructed from publicly available educational resources. NCERT textbooks are released under an open educational license by the Government of India, and JEE Mains examination papers are publicly distributed by the National Testing Agency. ChemPro will be released with full license compliance with all sources to ensure unrestricted access for research.

4. Analysis and discussion

4.1. Human performance comparison and evaluation

Our human performance reference draws from established performance data on JEE Mains and NCERT board examinations from 2020–2024, the same period from which ChemPro questions are sourced. Historical data from these examinations shows that top 100 students typically score 97%–100% (refer Fig. 7) on chemistry sections. We emphasize that this is an *indirect proxy*: the cited cohort did not take ChemPro end-to-end, and the exact ChemPro sampling may not match

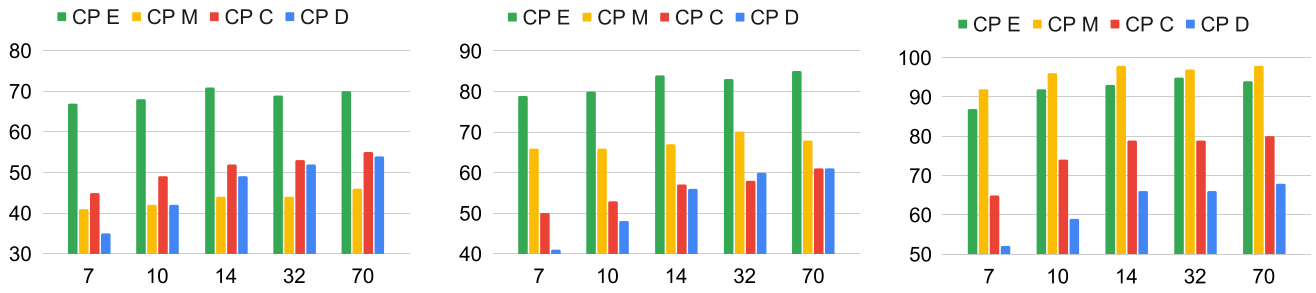

Fig. 4. ***Parameter scaling limitations.*** Scaling benefits plateau well below human performance levels (90%+ expected accuracy). Numerical reasoning limitations persist across all parameter scales (x-axis: Parameters in billions). **Left:** Average Model Performance on Numericals (y-axis: Exact Match score). **Center:** Average Model Performance on Numericals (y-axis: Tolerance-Based score). **Right:** Average Model Performance on MCQs (y-axis: Accuracy).

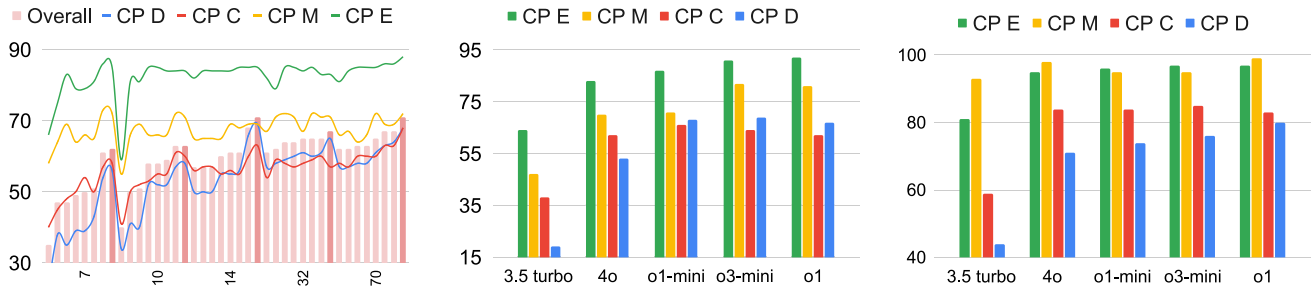

Fig. 5. ***Comprehensive performance analysis.*** **Left:** Overall tolerance-based performance across all model sizes showing scaling limitations. **Center:** OpenAI model performance on numerical problems across ChemPro sections (y-axis: Tolerance-Based score). **Right:** OpenAI model performance on multiple-choice questions across ChemPro sections (y-axis: Accuracy).

Table 3
Performance Comparison between agentic framework ChemCrow and GPT-4o on ChemPro MCQs.

System	CP_E	CP_M	CP_C	CP_D
ChemCrow	97	98	84	68
GPT-4o	95	98	84	71

any single historical paper. We therefore use this reference primarily as an educationally grounded upper-bound context for expected mastery of the underlying curriculum, rather than as a controlled human-vs-model head-to-head evaluation. This comparison is analogous to AlphaGeometry's [59] silver medal performance evaluation on International Mathematics Olympiad.

Significance: ChemPro tests whether broadly-deployed LLMs can handle chemistry material students are expected to master. In addition to reporting accuracy, we use the progressive design to diagnose where robustness breaks as problems require longer chains of operations (e.g., conversions and cross-condition integration). Table 3 further suggests that tool-augmented agentic frameworks alone do not remove degradation at higher tiers.

4.2. Empirical evidence for articulation effects

Complex articulation-defined by multi-step reasoning, unit conversions, and conceptual integration degrades performance across curriculum-aligned content. Importantly, we do not equate articulation complexity with surface linguistic length; rather, it reflects the structure of required operations (chained steps, conversions, and cross-condition integration).

ChemPro reveals this pattern (Figs. 4 & 5) through performance measurement across difficulty sections. The progression from CP_E to CP_D represents increasing difficulty within the same curricular scope, as evidenced by: 1. **Consistent Performance Degradation**: All 45 evaluated models show declining accuracy as section difficulty

Fig. 6. ***Dataset comparison:*** Performance of top models from each size category on ChemPro MCQs, College Chemistry (CC), and High School Chemistry (HSC). (x-axis: model sizes (7B to 70B & Proprietary); y-axis: accuracy).

increases, with an average ≈ 21-percentage-point drop from elementary to competitive formulations. 2. **Pattern Across Architectures**: This degradation pattern appears across diverse model architectures (gemma, Qwen, Falcon, PHI, etc.) [32,60–62] and scales (7B to 70B+), indicating a systematic challenge rather than architecture-specific limitations. 3. **Preserved Relative Rankings**: While absolute performance varies, the relative difficulty ordering remains consistent across models, confirming that articulation complexity represents a measurable but under-attended dimension of challenge. This effect is pronounced for numericals, where unit handling and multi-step calculations are frequent, and errors can cascade (see Fig. 6).

Articulation and Conceptual Complexity: Our curriculum-aligned design enables this distinction through educational provenance. The key insight is that CP_C and CP_D (JEE Mains) operate within identical curriculum boundaries, JEE Mains officially adheres to NCERT syllabus as defined by the National Testing Agency. Questions within each chemistry subdomain (biochemistry, organic, inorganic, physical

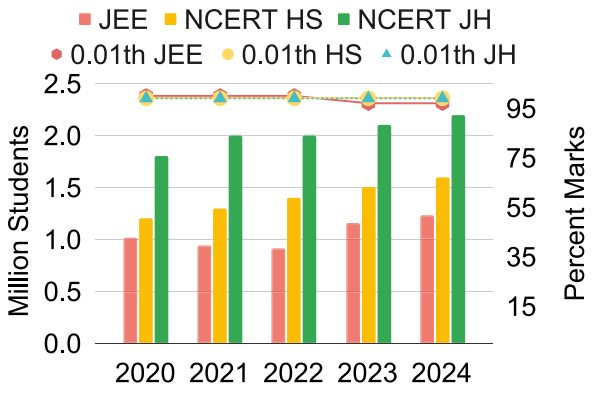

Fig. 7. ***Human vs. LLM performance comparison.*** Human performance on corresponding educational assessments significantly exceeds current LLM capabilities.

chemistry) therefore assess the same conceptual scope while varying in formulation complexity. The systematic nature of performance degradation between these curriculum-equivalent sections (13-point average drop), consistent across models, supports the interpretation that articulation patterns contribute to difficulty due to composition rather than missing topic coverage.

4.3. The open-source capability gap

Small Model ≈ 7B & 10B: Small models degrade sharply with competitive formulations, with numerical reasoning as a recurring failure mode. **Medium Model ≈ 14B & 32B**: Medium models improve on easy tiers but remain unreliable as formulation complexity increases. **Large Models ≈ 70B**: Even the strongest open models remain meaningfully below educational reliability, especially on numericals.

Overall, current open models do not reliably master foundational chemistry under competitive articulation.

4.4. Performance scaling analysis

Analysis shows that parameter scaling is not robust to academic complexity: performance drops substantially with ChemPro tiers, and numerical limitations persist even for the largest models. Scaling improves performance on easier tiers but exhibits diminishing returns as questions demand longer procedural chains, indicating that robustness to articulation complexity remains a key bottleneck.

5. Results

We structure our findings around four key observations:

Articulation as a Fundamental Challenge: Complex articulation represents a fundamental and systematic challenge for current LLMs, persisting across all tested model scales and architectures. Performance consistently degrades as questions progress from elementary formulations ($\mathcal{CP}_E$) to sophisticated multi-step reasoning requirements ($\mathcal{CP}_D$) within the same curricular scope. This pattern, marked by significant percentage-point performance drops, correlates with increased demands for multi-step reasoning, and conceptual integration, confirming that challenge lies in the reasoning process, with underlying scientific concepts.

Diminishing Returns with Model Scaling: While larger models outperform smaller variants on foundational tasks-consistently achieving $\geq$ 90% accuracy on $\mathcal{CP}_E$ MCQs-all models exhibit similar degradation patterns as formulation complexity increases (Fig. 4). We observe performance convergence in the intermediate $\mathcal{CP}_C$ and $\mathcal{CP}_M$ sections. Even top-performing proprietary systems (o1, o3-mini), which achieve 74%–76% accuracy on competitive questions, follow the same fundamental degradation pattern. For comparison, 7-10B models achieve 53%–67% on these questions, with 70B+ variants reaching 68%–71%. These patterns strongly suggest that parameter scaling alone is insufficient to overcome multi-step scientific reasoning.

Ineffectiveness of Current Agentic Frameworks: Our assessment of ChemCrow, reveals that current sophisticated reasoning frameworks do not overcome the identified limitations. ChemCrow provides only marginal improvements on elementary and intermediate tasks ($\mathcal{CP}_\mathcal{E}$ and $\mathcal{CP}_\mathcal{M}$) and fail to bridge the complexity gap for advanced problems ($\mathcal{CP}_C$ and $\mathcal{CP}_D$). For instance, it achieves performance comparable to its base GPT-4o model (Table 3), indicating that access to tools and prompting strategies does not fundamentally resolve the underlying reasoning bottlenecks.

Subfield-Specific Reasoning Bottlenecks: A detailed (Fig. 8) analysis reveals distinct and systematic bottlenecks across chemistry subfields. **Biochemistry** yields the highest accuracy but are often hindered by computational demands. **Organic Chemistry** shows severe performance drops on problems requiring advanced spatial reasoning and multi-step synthesis. In **Physical Chemistry**, models leverage mathematical formulations but frequently fail on precise numerical calculations and unit conversions. Finally, **Inorganic Chemistry** displays high variability, with unpredictable performance, indicating a fragile understanding of bonding, coordination chemistry and reactivity principles.

Our Evaluations on 7 additional (new) open-sourced models Table 7 (3 General Purpose and 4 Chemistry Focused) has been provided in Table 8(Supplementary). The scores from these models re-verify our inferences and thereby establish a strong need for **ChemPro** benchmark and research oriented towards robustness against complex articulation.

6. Conclusions

We introduce ChemPro, a novel curriculum-aligned progressive benchmark designed to rigorously evaluate and diagnose LLM capabilities in scientific reasoning. Our comprehensive evaluation of models

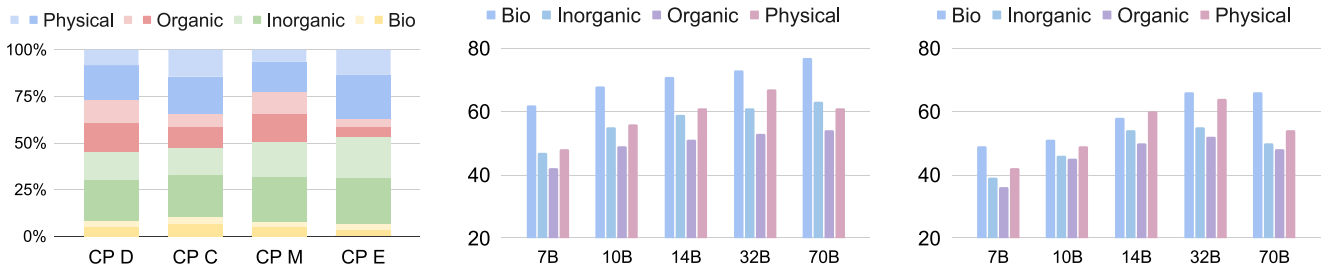

Fig. 8. ***Subfield performance analysis*** **Left:** Section-wise distribution of questions into subfields (y-axis: Dataset distribution; lighter shades are MCQs and darker shades are Numericals). **Center:** Model performance on multiple choice questions (y-axis: accuracy) across subfields. **Right:** Model performance on numericals (y-axis tolerance-based score) across subfields.

unequivocally demonstrates a systematic pattern of performance degradation directly correlating with question articulation complexity. This consistent decline, observed across all models regardless of architecture or scale, exposes fundamental challenges in current language modeling approaches to multi-step scientific reasoning.

The findings from ChemPro are critical: Current architectural paradigms face inherent limitations in complex scientific reasoning that cannot be overcome solely through parameter scaling or even sophisticated agentic frameworks. ChemPro's curriculum-aligned progression validates that these observed failures occur on material students are expected to master, confirming genuine reasoning limitations rather than a lack of esoteric expert knowledge. By systematically profiling reasoning depth and identifying subfield-specific bottlenecks, ChemPro serves as a powerful diagnostic tool and provides a clear path forward for developing next-generation LLMs capable of robust and reliable scientific problem-solving, beyond superficial understanding to true conceptual mastery essential for educational and research applications.

CRediT authorship contribution statement

Aaditya Baranwal: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shruti Vyas:** Writing – review & editing, Validation, Supervision, Project administration, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.aichem.2026.100118.

References

[1] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, 2024, arXiv:2402.06196.

[2] J. Jiang, F. Wang, J. Shen, S. Kim, S. Kim, A survey on large language models for code generation, 2024, arXiv:2406.00515.

[3] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, W. Yin, Large language models for mathematical reasoning: Progresses and challenges, 2024, arXiv:2402.00157.

[4] Y. Zhang, X. Chen, B. Jin, S. Wang, S. Ji, W. Wang, J. Han, A comprehensive survey of scientific large language models and their applications in scientific discovery, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 8783–8817.

[5] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? Try ARC, the AI2 reasoning challenge, 2018, arXiv:1803.05457.

[6] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022, arXiv:2209.09513.

[7] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021, arXiv:2009.03300.

[8] D. Rein, B.L. Hou, A.C. Stickland, J. Petty, R.Y. Pang, J. Dirani, J. Michael, S.R. Bowman, GPQA: A graduate-level google-proof q&a benchmark, 2023, arXiv:2311.12022.

[9] D. Arora, H.G. Singh, Mausam, Have LLMs advanced enough? A challenging problem solving benchmark for large language models, 2023, arXiv:2305.15074.

[10] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A.R. Loomba, S. Zhang, Y. Sun, W. Wang, SciBench: A college-level scientific problem-solving benchmark, 2023, arXiv:2307.10635.

[11] S. Yu, J. Zhou, X. Zhou, Y. Liu, J. He, Y. Kuroda, Y. Kawahara, H. Li, S. Liu, N. Cheng, T. Fu, Y. Zhu, J. Leskovec, R. Ying, SMolInstruct: A large-scale dataset and benchmark for structure-based molecular property prediction, 2024, arXiv:2406.13393.

[12] J. Ye, Y. Zhang, S. Zhang, L. Wang, J. Gong, L. Qi, J. Li, Y. Lin, J. Han, Z. Zhang, F. Wu, W. Liu, H. Zhou, J. Li, C. Huang, MolInstruct: A large-scale multi-task instruction tuning dataset for molecular understanding, 2024, arXiv:2306.08018.

[13] S. McNaughton, E. Elmoznino, M. Koziarski, F.L. Lopez, R. Luechtefeld, A. Sotiras, J. Clune, CACTUS: Chemistry agent benchmark for tool usage and synthesis, 2024, arXiv:2405.00972.

[14] X. Chen, T. Wang, T. Guo, K. Guo, J. Zhou, H. Li, M. Zhuge, J. Schmidhuber, X. Gao, X. Zhang, ScholarChemQA: Unveiling the power of language models in chemical research question answering, 2024, arXiv:2407.16931.

[15] A. Mirza, N. Alampara, S. Kunchapu, M. Rios-Garcia, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, A.M. Elahi, M. Asgari, J. Eberhardt, H.M. Elbeheiry, M.V. Gil, M. Greiner, C.T. Holick, C. Glaubitz, T. Hoffmann, A. Ibrahim, L.C. Klepsch, Y. Koster, F.A. Kreth, J. Meyer, S. Miret, J.M. Peschel, M. Ringleb, N. Roesner, J. Schreiber, U.S. Schubert, L.M. Stafast, D. Wonanke, M. Pieler, P. Schwaller, K.M. Jablonka, Are large language models superhuman chemists?, 2024, arXiv:2404.01475.

[16] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N.V. Chawla, O. Wiest, X. Zhang, What can large language models do in chemistry? A comprehensive benchmark on eight tasks, 2023, arXiv:2305.18365.

[17] P. Song, T. Kerber, M. Schilling-Wilhelmi, P. Friederich, M.V. Gil, J. Meiler, T. Siebert, P. Schwaller, K.M. Jablonka, RESTEEM: A data repository of educational materials for chemistry, 2024, arXiv:2406.04654.

[18] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, Galactica: A large language model for science, 2022, arXiv:2211.09085.

[19] T.L. Brown, H.E. LeMay, B.E. Bursten, C.J. Murphy, P.M. Woodward, Chemistry: The Central Science, fourteenth ed., Pearson, 2017.

[20] A. Agrawal, S. Gadgil, N. Goyal, A. Narayanan, A. Tadipatri, Towards a mathematics formalisation assistant using large language models, 2022, arXiv:2211.07524.

[21] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N.V. Chawla, O. Wiest, X. Zhang, What can large language models do in chemistry? A comprehensive benchmark on eight tasks, in: Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Datasets and Benchmarks Track, 2023.

[22] C. Liao, Y. Yu, Y. Mei, Y. Wei, From words to molecules: A survey of large language models in chemistry, 2024, arXiv:2402.01439.

[23] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021, arXiv:2110.14168.

[24] C.E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, K. Narasimhan, SWE-bench: Can language models resolve real-world GitHub issues?, 2024, arXiv:2310.06770.

[25] T. Trinh, Y. Wu, Q. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations, Nature (2024).

[26] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P.S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023, arXiv:2307.03109.

[27] National Testing Agency (NTA), JEE main official website, 2024, (Accessed 15 February 2024).

[28] National Council of Educational Research and Training (NCERT), NCERT official website, 2024, (Accessed 15 February 2024).

[29] Education Quizzes, Education quizzes website, 2024, (Accessed 15 February 2024).

[30] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL technical report, 2025, arXiv:2502.13923.

[31] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, G. Penedo, The falcon series of open language models, 2023, arXiv:2311.16867.

[32] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R.J. Hewett, M. Javaheripi, P. Kauffmann, J.R. Lee, Y.T. Lee, Y. Li, W. Liu, C.C.T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024, arXiv:2412.08905.

[33] T. Saeki, H. Zen, Z. Chen, N. Morioka, G. Wang, Y. Zhang, A. Bapna, A. Rosenberg, B. Ramabhadran, Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech, 2023, arXiv:2210.15447.

[34] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the MATH dataset, 2020, arXiv:2103.03874.

[35] M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P.d. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F.P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W.H. Guss, A. Nichol, A. Paino, N. Tezak,

J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A.N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, W. Zaremba, Evaluating large language models trained on code, 2021, arXiv:2107.03374.

[36] T. Madushanka, S. Rana, X. Zou, E. Bengtsson, A. Henriksson, S. Yuan, A. Adam, E.J. Schelter, Z. Shabbir, B. Nan, A. Sferruzzi, M. Ek, C. Kern, M. Ullrich, S. Strobel, H.J. Kulik, G.P. Wellawatte, P. Schwaller, O. Engkvist, G. Tom, ChemLactica: A large language model for chemistry, 2024, arXiv:2402.00746.

[37] J. Feng, L. He, X. Li, Y. Xu, Y. Li, Y. Zhou, Y. Cai, H. Zhang, K. Chen, Z. Tang, K. Qin, D. Yu, J. Li, C. Qin, X. Chen, M. Zhang, H. Lin, H. Li, J. Liu, J. Liu, Z. Zhang, G. Ke, Llama-chem: Large language model for chemistry, 2024, arXiv:2402.06852.

[38] D. Adak, Y.S. Rawat, S. Vyas, Molvision: molecular property prediction with vision language models, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025, https://openreview.net/forum?id=6vI3OOYddm.

[39] Suayptalha, HomerCreativeAnvita-Mix-Qw7B, 2025, (Accessed 02 August 2025).

[40] Suayptalha, Falcon3-Jessi-v0.4-7B-Slerp, 2025, (Accessed 02 August 2025).

[41] Tiiuae, Falcon3-7B-Instruct, 2025, (Accessed 02 August 2025).

[42] Zelk12, MT-Merge4-gemma-2-9B, 2025, (Accessed 02 August 2025).

[43] Tensopolis, falcon3-10b-tensopolis-v1, 2025, (Accessed 02 August 2025).

[44] Tiiuae, Falcon3-10B-Instruct, 2025, (Accessed 02 August 2025).

[45] Suayptalha, Lamarckvergence-14B, 2025, (Accessed 02 August 2025).

[46] Bunnycore, Phi-4-Model-Stock-v4, 2025, Accessed: 02 August 2025.

[47] Suayptalha, Luminis-phi-4, 2025, (Accessed 02 August 2025).

[48] Daemontatox, PathFinderAi3.0, 2025, (Accessed 02 August 2025).

[49] Zetasepic, Qwen2.5-32B-Instruct-abliterated-v2, 2025, (Accessed 02 August 2025).

[50] RomboOrg., Rombos LLM V2.5, 2024, (Accessed 15 February 2024).

[51] Shuttleai, shuttle-3, 2025, (Accessed 02 August 2025).

[52] Newsbang, Homer-v1.0-Qwen2.5-72B, 2025, (Accessed 02 August 2025).

[53] OpenAI, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A.T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C.M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F.P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H.W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I.C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J.Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M.Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R.G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S.R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, Z. Li, OpenAI o1 system card, 2024, https://arxiv.org/abs/2412.16720.

[54] OpenAI, OpenAI o3-mini System Card, 2025, Published: January 31, 2025; (Accessed 02 August 2025).

[55] OpenAI, A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mdry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A.T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C.J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D.P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F.P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H.P. de Oliveira Pinto, H. Ren, H. Chang, H.W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J.G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J.Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J.W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M.J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M.O.T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R.G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, Y. Malkov, GPT-4o system card, 2024, https://arxiv.org/abs/2410.21276.

[56] D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, Y. Li, ChemLLM: A chemical large language model, 2024, arXiv:2402.06852.

[57] Z. Zhao, D. Ma, L. Chen, L. Sun, Z. Li, Y. Xia, B. Chen, H. Xu, Z. Zhu, S. Zhu, S. Fan, G. Shen, K. Yu, X. Chen, Chemdfm: A large language foundation model for chemistry, 2024, arXiv:2401.14818.

[58] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025, arXiv:2505.09388.

[59] T. Trinh, Y. Wu, Q. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations, Nature (2024).

[60] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviere, M.S. Kale, J. Love, P. Tafti, L. Hussenot, P.G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Heliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C.L. Lan, C.A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway,

J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L.L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikula, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S.L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, K. Kenealy, Gemma: Open models based on gemini research and technology, 2024, arXiv:2403.08295.

[61] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025, arXiv:2412.15115.

[62] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, G. Penedo, The falcon series of open language models, 2023, arXiv:2311.16867.