JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access

This article is licensed under CC-BY-NC-ND 4.0

pubs.acs.org/jcim

Article

In-Context Molecular Property Prediction with LLMs: A Blinding Study on Memorization and Knowledge Conflicts

Published as part of Journal of Chemical Information and Modeling special issue "Harnessing the Power of Large Language Model-Based Chatbots for Scientific Discovery".

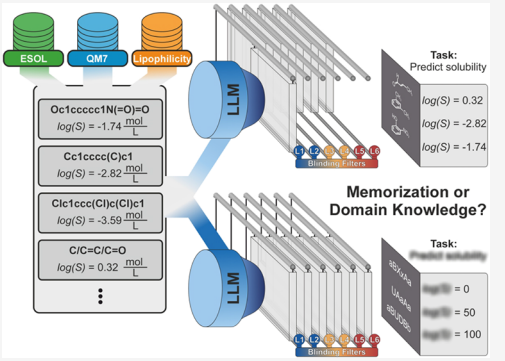
Matthias Busch,* Marius Tacke, Sviatlana V. Lamaka, Mikhail L. Zheludkevich, Christian J. Cyron, Christian Feiler,* and Roland C. Aydin*

Cite This: J. Chem. Inf. Model. 2026, 66, 12305−12315

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The capabilities of large language models (LLMs) have expanded beyond natural language processing to scientific prediction tasks, including molecular property prediction. However, their effectiveness in in-context learning remains ambiguous, particularly given the potential for training data contamination in widely used benchmarks. This paper investigates whether LLMs perform genuine in-context regression on molecular properties or instead rely on verbatim retrieval of memorized target values. Furthermore, we analyze the interplay between pretrained knowledge and in-context information through a series of progressively blinded experiments. We evaluate nine LLM variants across three families (GPT-4.1, GPT-5, Gemini 2.5) on three MoleculeNet data sets (Delaney solubility, Lipophilicity, QM7 atomization energy) using a systematic blinding approach that iteratively reduces available information, complemented by 0-, 60-, and 1000-shot in-context sample sizes as an additional control for information access. To validate the memorization analysis and the blinding experiments, we add a positive and a negative control for the memorization experiments and structural reference baselines for the multishot experiments, as well as bootstrap confidence intervals for all results. We find no evidence of verbatim retrieval on the legacy benchmarks and show that blinding exposes conflicts between pretrained knowledge and in-context information. This work provides a principled framework for evaluating molecular property prediction under controlled information access.



INTRODUCTION

Large language models (LLMs) have emerged as powerful tools across numerous scientific domains, driven by their ability to process natural language and perform few-shot learning from minimal examples.[1−3] In chemistry, LLMs have been explored for diverse tasks[4] including molecule generation[5] and property prediction.[6−8] The appeal of LLMs lies in their pretrained knowledge and in-context learning (ICL) capabilities, which enable predictions with limited task-specific training data—a critical advantage in domains where experimental data is scarce or expensive to obtain.

However, the application of LLMs to molecular property prediction presents fundamental questions about the nature of their predictive capabilities. Unlike classification tasks, prediction requires precise numerical reasoning and an understanding of continuous relationships between molecular structure and properties. Recent work by Joe et al.[8] has revealed that LLMs achieve near-perfect accuracy on raw molecular weight prediction from SMILES strings but fail when simple mathematical transformations are applied to the target values. This *molecular weight anomaly*[8] suggests that apparent success may stem from shortcut learning—exploiting superficial cues like token counting—rather than genuine structural reasoning.

Motivation: Data Set Memorization and Benchmark Contamination

A critical concern for evaluating LLMs on molecular property prediction is *data set contamination*: widely used benchmark data sets like ESOL[9] Lipophilicity, and QM7[10] have been publicly available for years and may appear in LLM training corpora. Recent studies have documented extensive memorization in LLMs[11,12] and raised concerns that high benchmark scores may reflect retrieval rather than learning.[13,14] When an LLM has encountered data set labels during pretraining, strong

Received: March 31, 2026
Revised: August 4, 2026
Accepted: August 6, 2026
Published: September 17, 2026



ACS Publications

© 2026 The Authors. Published by American Chemical Society

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854789/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

in-context performance does not necessarily indicate genuine pattern recognition—it may simply trigger recall of memorized values.

This contamination problem is particularly acute for molecular property prediction, where the number of widely used benchmarks is limited and repeatedly cited data sets (ESOL, MoleculeNet[15]) dominate the literature. While such data sets provide valuable standardized evaluation, their ubiquity makes them poor tests of an LLM's ability to learn new functional relationships from in-context examples. This motivates our central approach: *systematic information blinding* to test whether LLMs can generalize beyond memorization.

Research Approach: Blinding and In-Context Benchmarking

Inspired by the transformation framework of Joe et al.[8]—which used mathematical transformations to prevent shortcut learning on molecular weight—we introduce a complementary *six-level blinding approach* that progressively removes contextual information from the prediction task and transforms the target values. Our blinding levels range from full chemical context (Level 1: explicit property names, chemical terminology, untransformed values) to complete abstraction (Level 6: generic "sample property" prediction with transformed values and transformed SMILES strings, eliminating all domain context). This framework enables us to analyze the influence of different levels of prior knowledge on the molecular property prediction capabilities of LLMs.

As an initial investigation, we explore unblinded LLMs with varying amounts of in-context examples. We probe the models' prior knowledge via 0-shot prompting and test their in-context learning abilities with 60-shot and 1000-shot configurations. This experimental design serves dual purposes: (1) investigating potential conflicts between in-context sample information and prior knowledge, and (2) assessing the models' capability to extract structure−property relationships from large in-context data sets.

Research Questions

Our investigation addresses three key questions:

1. **Memorization vs domain knowledge**: To what extent do LLMs rely on data set memorization versus genuine domain knowledge or general prediction capabilities? How does this balance shift across different molecular properties and varying levels of information blinding?
2. **Interaction of prior knowledge**: How does an LLM's prior internal knowledge interact with provided in-context examples? Does strong prior knowledge facilitate learning, or does it hinder performance when it conflicts with the provided samples (e.g., in blinded settings)?
3. **Implications for benchmarking and practice**: What are the broader implications for the use of standard molecular data sets as benchmarks, given the risks of contamination? Most importantly, what does this mean for researchers aiming to deploy LLM-based approaches for practical molecular property prediction?

RELATED WORK

LLMs for Molecular Property Prediction

Large language models have been applied to chemistry for various tasks[4] including molecule generation[5] and property prediction.[6] While chemistry-specific models like ChemLLM[16] have been developed, our recent work Busch et al.[7] shows that general-purpose LLMs (GPT-4o, o1) with strong reasoning outperform domain-specialized models on regression tasks, motivating our focus on general LLM families.

For in-context learning (ICL) on regression, Akyürek et al.[17] demonstrate transformers can implicitly implement ridge regression. Critically, Joe et al.[8] reveal LLMs exploit shortcuts in molecular weight prediction, achieving near-perfect accuracy on raw values but failing under nonlinear transformations—motivating our combined blinding and transformation approach. Building on the former result, a fully blinded LLM can be viewed as performing automated featurization followed by in-context regression—a learned analogue of the classical recipe of mechanical featurization (e.g., bag-of-characters counts) followed by ridge regression.

Memorization, Contamination, and Context Effects

Carlini et al.[11] demonstrate LLMs verbatim reproduce training data, with memorization increasing with model size. Widely used benchmarks like ESOL[9] from MoleculeNet[15] are vulnerable to contamination.[17−19] Our blinding approach addresses this by removing contextual information and applying value transformations, complementing the mathematical transformation framework of Joe et al.[8].

Liu et al. 2024 show LLMs struggle with the "lost in the middle" phenomenon in long contexts. Our evaluation with up to 1000 examples tests whether LLMs extract functional relationships from large data sets.

Positioning

To our knowledge, this is the first work that systematically evaluates the in-context-learning (ICL) performance of LLMs for molecular property regression under controlled information removal (blinding). We further believe this is the first systematic study to probe LLM ICL performance on large *experimental* data sets with up to 1000 ICL samples, providing enough data to enable predictions even without prior knowledge. Earlier studies, such as Joe et al.[8], use orders of magnitude fewer samples for ICL and computed rather than experimental quantities. Jablonka et al.[6] 2024 and our own previous study, Busch et al.[7], also use fewer ICL samples. None of these works systematically removed information from the context window to analyze its influence. In this work, we analyze which capabilities LLMs rely on when learning in-context and predicting molecular properties. For this, we use models from three LLM families across three popular benchmark data sets. We chose the data sets such that we expect the LLMs to know them because they all predate 2018. We do this because we want to also see memorization effects if they do exist. To put these effects in context, we additionally use control data sets: one created after the models' knowledge cutoff date, and two small data sets drawn from sources guaranteed to be in the training corpus.

METHODOLOGY

Data Sets

We evaluate on three MoleculeNet benchmark data sets:[15] ESOL[9] measures aqueous solubility for 1,128 organic compounds (log mol/L), Lipophilicity contains octanol−water partition coefficients (logD at pH 7.4) for 4,200 compounds, and QM7[10,20] provides DFT-computed atomization energies (kcal/mol) for 6,834 small molecules. We emphasize that for ESOL we use the *experimentally measured* log-solubility column as the target; the data set's *calculated*

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

Table 1. Six-Level Blinding Framework[a]

Level	Name	Label	Input	Context description
1	Specific	Original	Original	Full chemical context: property name (e.g., "log solubility"). Tests memorization + domain knowledge
2	Specific-Transformed	Transformed	Original	
3	Generic	Original	Original	Generic "molecular property" label. Tests chemical in-context learning without property-specific priors
4	Generic-Transformed	Transformed	Original	
5	Agnostic	Original	Transformed	Domain-agnostic framing ("sample property"). Tests general in-context learning
6	Agnostic-Transformed	Transformed	Transformed	

[a]Each level progressively reduces contextual information to isolate different LLM capabilities. "Label" indicates if target values are mathematically transformed; "Input" indicates if SMILES strings are transformed.

ESOL value (the output of Delaney's regression equation) is used neither as a target nor as an input feature. Likewise, the Lipophilicity target is the *experimental* logD at pH 7.4; the model only ever receives the SMILES string, never a computed descriptor. Note that we use a HuggingFace variant of QM7 with SMILES strings as sole descriptors rather than the standard version with 3D coordinates, resulting in slightly fewer samples; as a consequence its structure−property statistics differ from the canonical 3D QM7 (see SI). These data sets provide complementary challenges: experimentally measured physicochemical properties (ESOL, Lipophilicity) versus quantum-mechanical predictions (QM7), with varying data set sizes and chemical diversity. In all cases, the only input provided to the model is the SMILES string[21] of the molecule, together with the target property value.

Additionally, we used a set of 41 IUPAC atomic masses and 45 boiling points reported on Wikipedia as a positive memorization control, and the 2025 ASAP Discovery antiviral potency data set from the ASAP × Polaris × OpenADMET blind challenge[22] as a negative control.

Six-Level Blinding Framework

To test whether LLMs perform genuine in-context learning versus memorization or shortcut exploitation, we introduce a systematic *information blinding* approach that progressively reduces contextual information. Our framework comprises six levels (Table 1).

The "specific" levels specify the molecular property (e.g., "exact" = "log solubility" or "close" = "property related to solubility"), allowing LLMs to apply domain knowledge and use memorized values. Levels 3−4 use generic "molecular property" labels, preventing domain-specific reasoning while retaining chemistry terminology. Levels 5−6 remove all chemistry context ("samples", "structure strings" instead of "molecules", "SMILES") and transform SMILES strings (see Section), forcing pattern recognition on an unfamiliar structural language. Levels 2, 4, and 6 apply the transformation described in Section to reduce the impact of memorized values. Detailed prompts and blinding descriptions are in Supporting Information.

To interpret our findings and their implications, we introduce a framework distinguishing four levels of LLM capabilities in molecular property prediction. Each level corresponds to a specific level of generalization. With blinding the data set and the prompt, one can deactivate certain capabilities, and by comparing the results of different blinding levels, one can qualitatively analyze the influence of the different capabilities on the overall performance.

1 **Direct memorization**: The ability to exactly reproduce a value from training data without generalization. (Used by blinding level 1, and possibly 3 and 5).
2 **Learned structure−property relationships**: The ability to approximate values for known properties using learned correlations from specified input data like SMILES strings. (Used by blinding levels 1−2).
3 **Chemical In-Context Learning**: The ability to predict unknown properties for a given molecular structure by applying general chemical principles and extracting structure−property relations from molecular in-context examples. (Used by blinding levels 1−4).
4 **General In-Context Learning**: The ability to predict unknown properties solely from unknown structural patterns (e.g., SMILES syntax, but transformed) and in-context examples, without relying on chemical priors or property labels. (Used by blinding levels 1−6).

This taxonomy allows us to dissect the "performance" of LLMs not as a monolithic metric, but as a composite of these distinct mechanisms. It is important to note that these capabilities are not mutually exclusive, nor do they combine linearly, and experiments have to be designed carefully to disentangle them. In addition to the blinding levels, using 0-shot deactivates capabilities 3 and 4 since they rely on generalization from in-context examples while capabilities 1 and 2 rely only on the model's pretrained knowledge and the provided input data.

Transformation Functions

For blinding levels 2, 4, and 6, we apply a single mathematical transformation to the target values, which inverts the labels (multiplication by −1) and rescales them to the range [0, 100]. The inversion disrupts value memorization, while the rescaling prevents simple recognition of the original scale; both still preserve the qualitative relationship between input and label. This single transformation is applied consistently across all three blinded levels and all data sets. We report all metrics on inverse-transformed predictions to enable fair comparison across blinding levels. The benefit of a linear transformation is that it keeps the base difficulty of prediction exactly the same while preventing direct verbatim recognition of the values in the data set. As a control experiment, we also ran a nonmonotonic double-sine transformation, which led to similar results (see SI).

For blinding levels 5 and 6, we additionally transform the SMILES strings themselves to prevent the LLM from recognizing familiar chemical structures. We apply a deterministic character substitution where each atom symbol or SMILES syntax character is replaced by a unique substitute character. For example, atoms like C, N, O, Cl are mapped to arbitrary characters, and SMILES syntax elements like parentheses, brackets, and bond symbols are similarly replaced. This transformation is applied consistently across all molecules in both training and test sets. Critically, this substitution

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

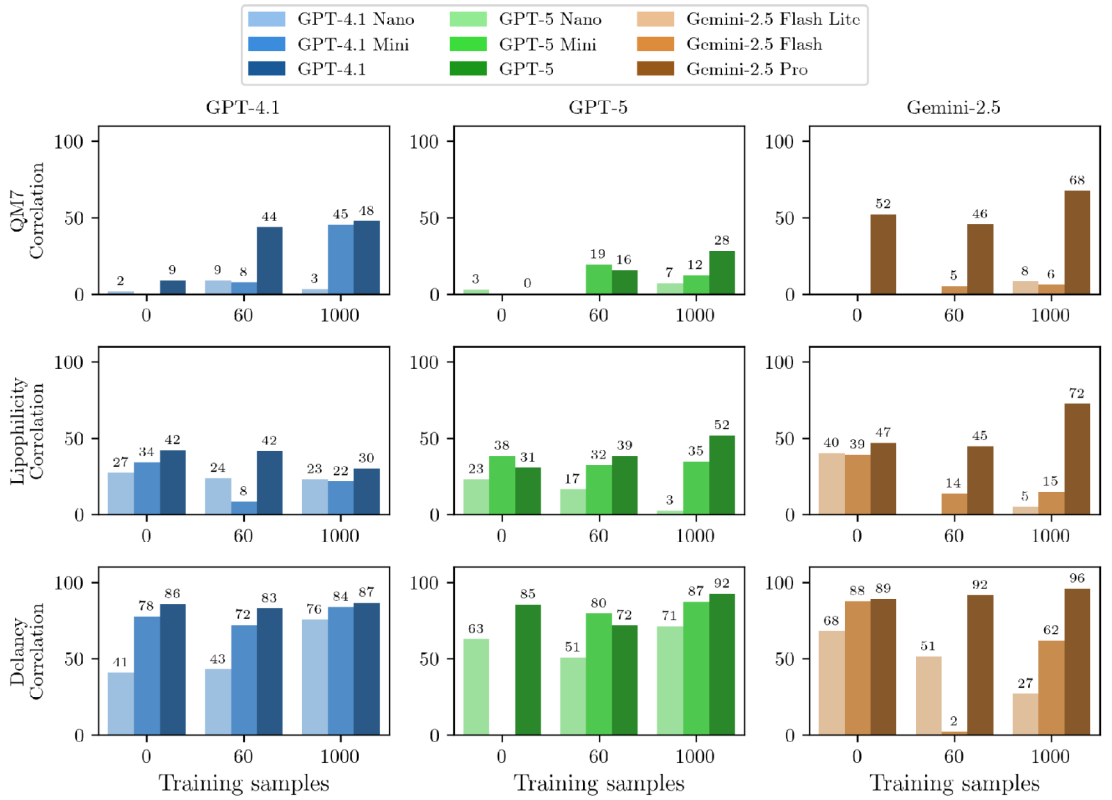


Figure 1. Correlation coefficients for molecular property prediction across three data sets (rows: QM7, Lipophilicity, Delaney) and three LLM families (columns: GPT-4.1, GPT-5, Gemini 2.5). The *x*-axis represents the input configuration (0-shot, 60-shot, 1000-shot). Different colors indicate model sizes within each family. Higher correlation values indicate better performance. Note the varying 0-shot correlations across the different data sets, indicating the popularity of the data sets in the training corpus, while they are relatively constant across the different model families. Also note that the 1000-shot correlation is usually higher than the 0-shot correlation, while the 60-shot correlation varies.

preserves the structural relationships encoded in the SMILES syntax—branching, ring closures, and connectivity remain intact—while rendering the strings unrecognizable to the LLM. This approach is analogous to the linear transformation of labels: it maintains the structure−property relationships that can be learned from in-context examples while disabling direct lookup of memorized chemical knowledge.

Prompting Strategy

We employ the preanalysis prompting approach of Busch et al.[7], which separates prediction into two phases: (1) Analysis: systematically examine training data to identify structural features, patterns, and similar molecule pairs. (2) Prediction: find similar training molecules to the test case, rank by similarity, and compute a weighted average prediction. This structured approach encourages chain-of-thought reasoning.[23] Furthermore, for the 0-shot experiments, we apply a simple prompt that asks the model to predict the target value without reasoning. Details of the prompts are provided in Supporting Information.

In-Context Learning Setup

We evaluate three configurations: 0-shot (no training examples), 60-shot (60 random training molecules), and 1000-shot. We use the same random train-test split with 150 test samples for all runs except the 0-shot runs (1000 test samples) to balance statistical robustness. We perform 2 runs per configuration and data set to limit the computational cost.

Evaluation Metrics

We evaluate prediction quality using the Pearson correlation coefficient (r), since MAE and RMSE would be affected by the transformation's scale shift and by the differing scales of the data set labels, whereas the correlation is invariant under such rescaling. Only the cumulative error distributions use MAE as the single-sample error metric. All metrics are computed on a held-out test set not included in in-context training examples.

Implementation Details and Large Language Models

We evaluate nine LLM variants across three families and three size classes: GPT-4.1 (`gpt-4.1-nano`, `gpt-4.1-mini`, `gpt-4.1`) and GPT-5 (`gpt-5-nano`, `gpt-5-mini`, `gpt-5`) by OpenAI, and Gemini 2.5 (`gemini-2.5-flash-lite`, `gemini-2.5-flash`, `gemini-2.5-pro`) by Google. This 3 × 3 design enables comparison of both family specific capabilities and size-dependent scaling effects. The LLMs are accessed via Azure (GPT models) and OpenRouter (Gemini models) with sampling parameters (temperature = 0.7, top_p = 0.95), where available. GPT-5 family models do not allow setting the temperature and top_p parameters.

RESULTS

General Performance and In-Context Learning Scaling

Figure 1 displays correlation coefficients across three LLM families, three data sets, and three input configurations (0, 60, and 1000 shots) without blinding. While the largest models generally perform well given that they only receive SMILES

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854789/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

strings, they do not surpass published baselines like Chemprop[24] or MPNNs.[25]

However, the purpose of these experiments is to check the pretrained capabilities (memorization and learned relationships) with the 0-shot experiments and possible improvements with in-context learning. The 0-shot performance is mixed for the different data sets but relatively consistent over the different models (with some exceptions). For QM7, the 0-shot performance is around 0 for all models except Gemini 2.5 Pro with around 53%, for Lipophilicity, the 0-shot performance is around 30−40% for all models and for Delaney, the 0-shot performance is around 85−90% for the largest models and decreases for smaller models. Since the 0-shot performance measures the pretrained capabilities and completely ignores in-context learning, the results are a proxy for the prevalence of the respective data set or the general underlying property in the models' training corpus. Hence, our results indicate that the Delaney data set or solubility is the most popular among the three, followed by Lipophilicity and then QM7 being seemingly absent from the models' effective training data. It should be noted that the variant of QM7 we use here provides SMILES strings rather than the standard 3D representation, which might also be a reason why most models do not recognize it.

When adding samples and in-context learning, the performance changes differently for the different models. The largest models behave similarly for all data sets: their performance increases in 8/9 cases from 0 to 1000 shots. However, the performance in 60 shots is in about half of the cases worse than the 0-shot performance. This indicates that the models struggle to combine their internal knowledge with a small number of samples but improve once a critical mass of samples is provided. Although individual per-model differences fall within the test-set confidence interval, these directional trends are jointly significant across the nine models by a sign test (1000-shot above 0-shot for the flagship models, $p = 0.02$; 60-shot below 0-shot, $p = 0.026$; 1000-shot above 60-shot, $p = 0.003$; Supporting Information, significance analysis). Middle and small-sized models show inconsistent performance. Gemini Flash and Flash-Lite clearly decrease or stagnate near 0 with more samples. OpenAI's mini and nano models are more robust; while their 0-shot average is lower, they handle additional samples effectively, maintaining performance without the degradation seen in the Gemini models.

Error Distribution and Memorization Analysis

Figure 2 illustrates the cumulative error distribution for 0-shot predictions on the Delaney data set. As can be seen, there is a continuous distribution of errors without a substantial amount of errors at zero. If the models had memorized the target values, we would expect a significant fraction of samples with exactly zero error. Since this is not the case, we conclude that the models do not reproduce the target values verbatim. This is consistent with the models relying on a learned relationship between SMILES strings (or molecular structures) and solubility, though it does not by itself exclude approximate recall or other statistical associations acquired during pretraining (see the Discussion).

To quantify potential memorization effects in greater detail, we analyzed the exact memorization of sample properties by the LLMs for the three legacy data sets we use throughout our study: Delaney, Lipophilicity, and QM7. Additionally, we included three control data sets: on the one hand a post

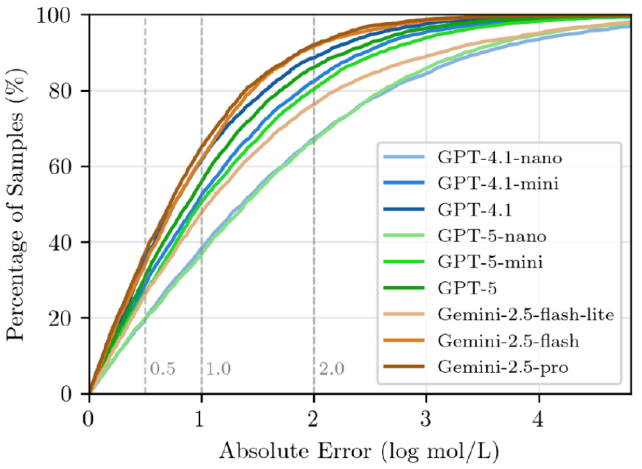


Figure 2. Cumulative error distribution for 0-shot predictions on the Delaney data set. The *x*-axis shows the absolute error threshold, and the *y*-axis shows the percentage of samples with error below that threshold. Steeper curves indicate better performance. Note the continuous distribution of errors without a substantial number of errors at zero, which would be the expected behavior if the LLMs had memorized the target values.

knowledge cutoff data set as negative control for retrieval—the 2025 ASAP Discovery antiviral potency data set—and on the other hand two additional data sets as positive control where direct verbatim retrieval is very likely—a collection of 41 IUPAC standard atomic weights and a set of 45 boiling points from Wikipedia.

For all 6 data sets, we report the exact match rate (first three significant digits), two precision-scaling *retention* ratios (m_2/m_1 and m_3/m_2) that measure the fraction of matches at one precision that survive to the next significant digit, and the median Pearson correlation of the per-molecule mean prediction. Because a value reported to one or two digits can match trivially, all match rates are computed over ground-truth values carrying at least three significant digits. A bare match rate, however, only catches *lossless* recall; Table 2 therefore pairs it with a precision-scaling retention statistic.

Table 2 shows the results of our 0-shot experiments, pooled across all models (detailed results in the SI). One can clearly see three times the same pattern. First, looking at the bare 3-significant-digit match rates, the results are split between the positive control data sets and the negative control. The models can recall IUPAC atomic weights almost with 100% certainty and Wikipedia boiling points mostly above 50% (see SI). On the other side sit the Antiviral potency data set as negative control with near zero match rate (0.13%). The three legacy benchmarks are similar in match rate to the negative control, with near-zero match rates (0.16−0.33%). Second, the precision-scaling retention ratios show a similar split: The positive controls are close to 100% retention, while the three legacy benchmarks and the postcutoff antiviral control are all around the ∼10% coincidence floor at *both* precision steps. Third, the median Pearson correlation shows this split as well, but on a much more continuous scale, with the three legacy benchmarks and the negative control showing a range of 0.01−0.78 and the positive controls showing near-perfect correlation (0.88−1.00).

First of all, the positive control shows how LLMs perform on a data set where they have demonstrably memorized the values. The results show large fractions of 3-significant digit

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

Table 2. Memorization Diagnostics (0-Shot, Pooled over the Nine Models)[a]

Data set	3-Sig match %	m_2/m_1	m_3/m_2	Median r
Delaney (ESOL)	0.33	10.8%	12.2%	0.78
Lipophilicity	0.33	11.6%	11.0%	0.39
QM7	0.16	17.8%	7.6%	0.01
Antiviral potency (postcutoff)	0.13	9.3%	11.3%	0.37
Atomic weights (memorized)	95−100	99.9%	97.7%	1.00
Boiling points (memorized)	16−78	81.7%	83.8%	0.88

[a]For each dataset wereport the three-significant-digit exact-match rate; the two precision-scaling *retention* ratios m_2/m_1 and m_3/m_2—the fraction of matches at one precision that survive to the next significant digit, with coincidence floor ≈ 10% and genuine recall ≈ 100% (the lower-precision m_2/m_1 rests on roughly 10× more matches and is the statistically tighter estimate); and the median Pearson r of the per-molecule mean prediction. Note that the three legacy benchmarks' results match qualitatively with the 2025 post-cutoff antiviral control, all sitting at the coincidence floor on *both* ratios together with near-zero match rates, whereas the positive control of demonstrably memorized values (IUPAC standard atomic weights; English-Wikipedia boiling points) is near-total on both under the *identical* detector. Detailed per-model counts, the full per-model retention tables, and the positive control are in the Supporting Information.

matches, with near-total retention at both precision steps. On the other side, the negative control of the 2025 postcutoff antiviral potency data set shows that even on an unseen data set, the models can achieve a nontrivial correlation (0.37) and a minimal exact-match rate (0.13%). However, the retention ratios show clearly that the correlation is based on the first significant digit, with the second and third digits being essentially randomly distributed. With one of ten possible digits matching by chance, i.e., ≈10% (for QM7 the second digit is not perfectly evenly distributed; see SI). The three legacy benchmarks show the same pattern as the negative control, which is why we conclude that the models probably do not retrieve the values of these specific data sets verbatim. While correct prediction of a single digit in 40−50% of the cases could theoretically be attributed to lossy recall of a memorized value, this rather looks like a very coarse approximation of the underlying chemistry, with the correct order of magnitude and leading digit being inferred from the structure.

We stress that this analysis does not prove that the data sets are absent from training (that would require access to the undisclosed corpora), but it provides strong evidence that the models have at least not memorized large fractions of their values during training. It also does not rule out approximate (lossy) recall, substructure-level associations, or other statistical regularities absorbed during pretraining, which may still contribute to the 0-shot predictions. But although the three legacy data sets clearly predate the development of the LLMs, they show the same characteristics as the data set that was certainly not part of the training corpus. The 0-shot correlation instead reflects how strongly the underlying property was represented in the training corpus.

As a test of whether our blinding protocols can suppress recall, we applied them to the positive control data sets, albeit with a different prompting strategy (see SI for full details). The results show that the stronger LLMs can uncover an inverted and rescaled property from in-context examples when they have memorized the underlying property and are given the property name. They can also find the underlying property if the name is masked. However, combining both the masking of the property name and the transformation of the target values suppresses a large part of the recall, but not completely. The agnostic blinding levels, which additionally transform the structure representation, bring recall down to the coincidence floor: retention there is statistically indistinguishable from chance, although the point estimates leave room for a small residual leakage that 45 compounds cannot resolve. The blinding is therefore effective at suppressing recall even on a data set the models have clearly memorized, but only in its fully blinded form: the property name alone protects nothing, and the value transformation protects only against the weaker models.

Effect of Blinding and Information Removal

Figures 3 and 4 detail the performance of the Gemini 2.5 and OpenAI (GPT-4.1 and GPT-5) model families, respectively, under progressive information removal (Levels 1−6) with in-context learning and 1000 samples. By the systematic removal or transformation of information, we can estimate the influence

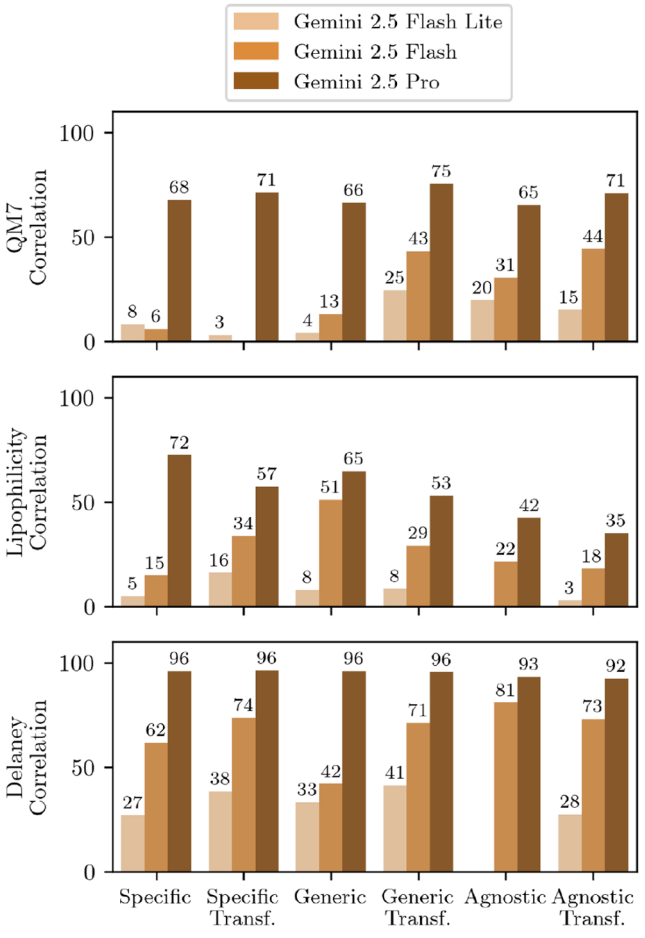


Figure 3. Detailed performance breakdown for Gemini 2.5 models across all six blinding levels (x-axis) and three data sets (rows). Specific blinding levels specify the property to be predicted by name, while generic blinding levels name it as molecular property and agnostic blinding levels name it as sample property. "Transf." relates to the label values being transformed as described in the Methods section. Note that the peak correlation value is frequently not at the first blinding level or approximately shared across different blinding levels.

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854789/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

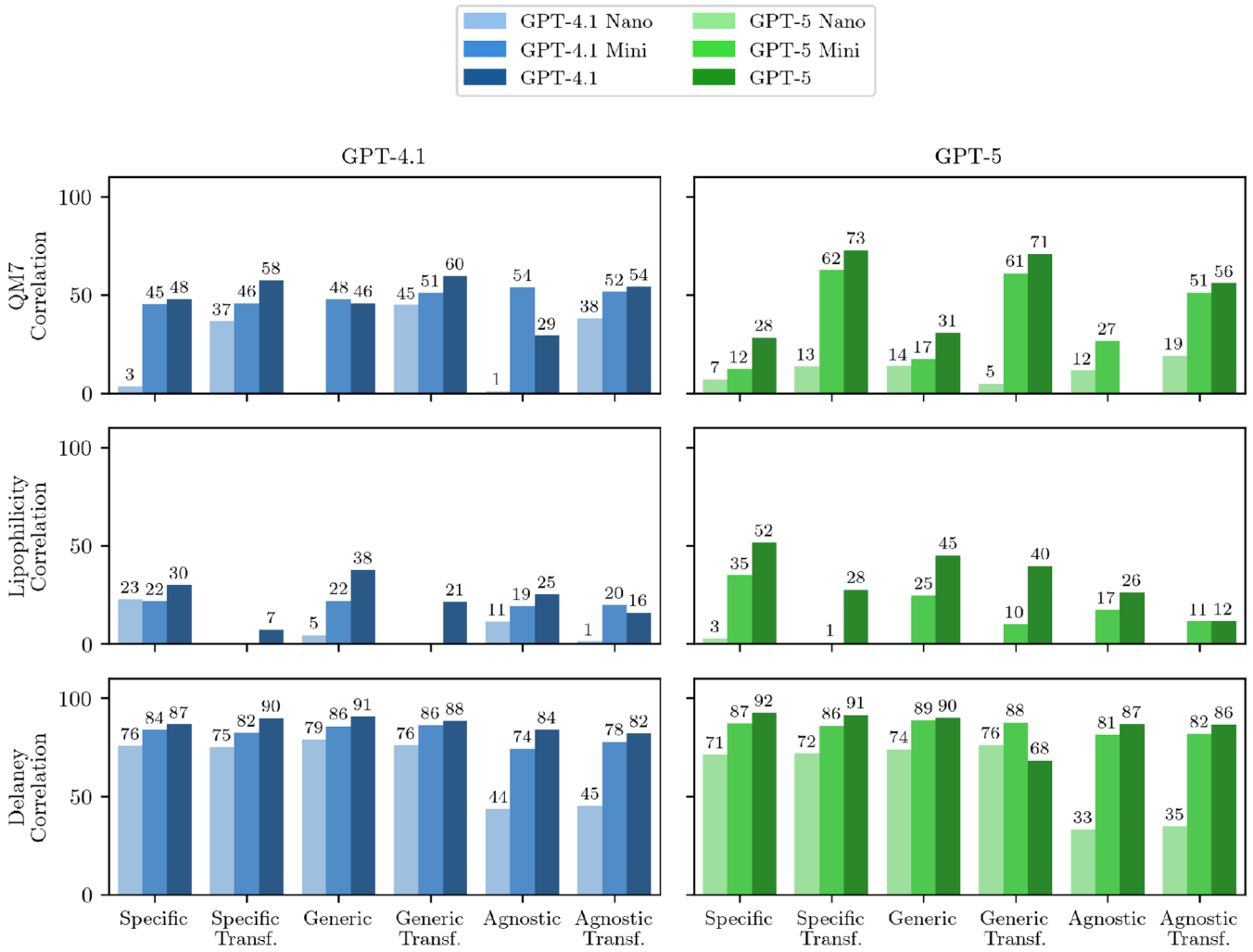


Figure 4. Correlation of OpenAI models (GPT-4.1 and GPT-5 families) across all six blinding levels and three data sets (rows: QM7, Lipophilicity, Delaney). Different colors indicate model sizes within each family. Compared to the Gemini results in Figure 3, the OpenAI models exhibit more uniform behavior across size variants, and the two families differ markedly in which blinding regime yields peak performance.

and strength of different capabilities of the LLMs. Additionally, we can compare the underlying difficulty levels of the prediction tasks by comparing the LLM performance on the fully blinded tasks. To confirm that the transformed levels probe genuine in-context learning rather than recovery of the original value scale from a monotonic map, we additionally validated them with a nonmonotonic sine transform that cannot be recovered from in-context anchors; across GPT-5, GPT-4.1, and Gemini 2.5 Pro the models retain their in-context performance under it—above both the structural-similarity and recall baselines—at every transformed level (nonmonotonic transform control, Supporting Information). Throughout this section we emphasize that, given only two runs per configuration, only differences whose bootstrap confidence intervals separate (Supporting Information) should be read as robust; the smaller, few-percentage-point gaps we describe are reported as qualitative trends rather than established effects.

Gemini 2.5 Pro consistently performs best but reveals distinct sensitivities. On Lipophilicity and Delaney, performance decreases with blinding, but the drop is much steeper for Lipophilicity (72% → 35%) than for Delaney (96% → 92%). This significantly larger drop indicates substantial reliance on pretrained knowledge (rather than verbatim memorization) for Lipophilicity, together with relatively weak in-context learning performance. In particular, a modest 5−15 percentage-point gap between transformed and untransformed values may reflect either a value-range effect or a memorization effect in which the model recognizes the original data set values; this gap is comparable to the test-set confidence interval, so we do not overinterpret it, and—given our negative memorization finding—we tentatively attribute it to a value-range issue rather than to recall. The robustness on Delaney even under full blinding suggests both that the structure−property relationship of solubility is not too complex and that the model can learn this relationship almost completely via in-context learning.

QM7 reveals a different behavior: there is no clear decrease in performance with blinding. However, there is an oscillation between transformed and nontransformed values, but such that the transformed values are predicted with higher correlation. This might just be due to the value range of the original values (around −1000) being highly suboptimal for LLM processing. Also, together with the fact that the performance is not decreasing with the blinding, it suggests that the model has neither memorized the data set nor has it learned the

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854789/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

relationships and is using its general in-context learning capability mostly for prediction.

Gemini Flash exhibits anomalous "interference" behavior, often performing better at higher blinding levels on difficult data sets. On QM7, masking the property name (agnostic) improves performance. On Lipophilicity, it peaks at Level 3. We hypothesize that for less familiar data sets, the model possesses "wrong knowledge" or hallucinations; blinding silences these incorrect priors, allowing the model to rely more effectively on the ground-truth in-context examples. Flash-Lite generally shows low performance, partly following Flash's trends on QM7 but struggling with larger data sets like Delaney.

While many trends are similar to the Gemini models, the OpenAI models show some distinct patterns. First, the OpenAI mini and nano variants perform much more similarly to their larger variants. This contrasts with the Gemini models, where the Flash and Flash-Lite variants show very different patterns compared to Gemini 2.5 Pro. Second, the GPT-4.1 models tend to peak with the Generic approaches and the GPT-5 models with the Specific approaches; because these peak differences are only a few percentage points and are not individually CI-separated, we report them as trends only. Taken together with the observation that both families reach similar correlation scores on the fully blinded Agnostic approaches, they are suggestive of GPT-5 combining prior knowledge and in-context samples somewhat more effectively than GPT-4.1, but this should not be read as a robust per-model claim.

The sole exception is the QM7 data set with original values, where GPT-5 achieves much lower correlation scores. On this data set, GPT-5 and GPT-5 mini show a similar oscillatory behavior, but with much larger amplitudes between the transformed and original values. This oscillatory behavior is also observed with Gemini 2.5 Pro and GPT-4.1, but on a much lower amplitude. This extreme sensitivity suggests that the large, negative original numbers in the QM7 data set (approximately −500 to −2500 kcal/mol) confuse the numerical reasoning of the GPT-5 models.

DISCUSSION

Key Findings and Implications

Our systematic evaluation across nine LLM variants, three data sets, and six blinding levels reveals several insights into how LLMs balance prior knowledge and in-context learning data when performing molecular property prediction.

Memorization vs Genuine Reasoning (Addressing RQ1)

Our results demonstrate that LLMs rely on different capabilities when performing molecular property prediction. Crucially, the 0-shot performance on a data set published after the models' knowledge cutoff shows that LLMs do generalize to unseen samples, consistent with their having learned a general structure−property relationship rather than retrieving specific values. The behavior on the data sets predating 2018 is very similar, indicating that predictions there, too, are not driven by direct, verbatim retrieval. We caution, however, that ruling out verbatim retrieval does not establish that the predictions arise purely from learned reasoning: approximate recall, substructure-level associations, or other statistical signals acquired during pretraining may also contribute to the observed 0-shot correlation. When given access to in-context examples, they additionally leverage in-context learning capabilities, as the increase in performance from 0-shot to 1000-shot indicates. However, a control test with IUPAC atomic weights and boiling points from Wikipedia showed that LLMs do retrieve values verbatim for numbers that are present in their training corpus.

When prediction must rely on in-context learning (enforced via blinding), performance varies substantially across data sets and model families. Lipophilicity is an example of a property that LLMs appear to recognize conceptually but struggle to learn via in-context examples, resulting in a significant performance drop from approximately 72% to 35% under blinding (consistent across 8/9 models, sign-test $p = 0.02$, and individually CI-separated for the strongest models; Supporting Information, significance analysis). Solubility, in contrast, shows that LLMs are able to learn the mapping in context, with only a minor drop from approximately 96% to 92%. The QM7 results even indicate no performance degradation under blinding. These differences likely reflect the underlying complexity of the structure−property relationships: Lipophilicity depends more strongly on global molecular structure, whereas solubility (Delaney) and atomization energy (QM7) are more directly associated with functional groups or atomic properties and hence can be learned more easily via in-context learning.

This progression also invites a mechanistic reading. Under full blinding the model receives only transformed structure strings and in-context examples, yet still predicts well above chance—in effect performing automated featurization followed by in-context regression (cf. Akyürek et al.[17]). A ridge regressor on a bag-of-characters representation of the SMILES string fit on 1000 examples (char-1000; Supporting Information, Table S8) is never clearly outperformed by an LLM on a fully blinded approach. What the models add from their knowledge and other capabilities becomes visible only as chemical context is restored through unblinding. However, exactly this additional knowledge can also interfere with in-context learning, as we discuss in the next section.

Taken together, these findings suggest that LLMs use their learned structure−property relationships if possible, but are also able to utilize chemical in-context learning to improve over the 0-shot setting. They even use general in-context learning in the absence of direct chemical context, although the complexity of the property to predict limits the performance of their in-context learning.

The Double-Edged Sword of Prior Knowledge (Addressing RQ2)

In-context learning is often framed as a process where prior knowledge facilitates new learning. Our findings refine this view: prior knowledge is a double-edged sword that can facilitate or *interfere* with learning, depending on its alignment with the specific task. We observe that the peak performance is often not at blinding level 1, consistent with prior knowledge hindering in-context learning. Most individual per-level differences are within the test-set confidence interval, so we do not overinterpret them; in specific cases, however, the effect is large and statistically significant—most clearly Gemini 2.5 Flash on QM7 (discussed below), where suppressing the property prior raises the correlation by roughly +38 percentage points (paired-bootstrap confidence interval excluding zero). Averaged across models the property name alone is not decisive—naming it generically (Level 3) is statistically indistinguishable from naming it specifically (Level 1)—and

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6845789/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

the most robust pooled signatures of interference are the few-shot dip discussed below and the Lipophilicity collapse (Supporting Information, significance analysis).

Furthermore, our results show a decrease in performance when using 60 samples instead of 0 in 19 of 27 model × data set cases (sign-test $p = 0.026$), even when the models are given reasoning time. For all models apart from Gemini 2.5 Flash and Flash-Lite, this effect is mostly mitigated by providing 1000 samples (1000-shot above 60-shot in 21/27 cases, $p = 0.003$; Supporting Information, significance analysis). This suggests a conflict between insufficient in-context data and the model's strong prior knowledge, where too few examples fail to override preexisting incorrect assumptions about the task. This is mostly resolved when providing 1000 samples.

Gemini 2.5 Flash and Flash-Lite show a stronger decrease in performance, leading to a lower performance with 1000 samples than in the 0-shot setting. Interestingly, this decrease is at least partly removed upon increasing the blinding level. This indicates that these two models cannot combine their knowledge with the in-context examples. As soon as we deactivate the models' ability to use its knowledge, there is only one source of information and the models' performance increases again. On the QM7 data set, this interaction produces particularly counterintuitive behavior: the models perform poorly without blinding, but exhibit sharp performance gains once the property name is hidden and inputs or outputs are transformed. For Gemini 2.5 Flash this gain is large and statistically significant—the QM7 correlation rises from approximately 6% at Level 1 to approximately 44% at Level 6, a +38 percentage-point improvement whose paired-bootstrap confidence interval [+17, +54] excludes zero (Gemini 2.5 Flash-Lite trends the same way, +15 pp, within noise). This suggests that these models rely on misleading prior associations for QM7-like tasks, which are only suppressed once prior knowledge is neutralized through blinding.

In conclusion, we find evidence that prior knowledge can be beneficial, but is often detrimental, especially when the amount of in-context examples is limited. This suggests that practical LLM deployment should carefully consider whether to leverage or suppress domain priors based on prior knowledge, data and LLM capabilities.

Implications for Benchmarking and Practice (Addressing RQ3)

When benchmarking LLMs on molecular property prediction, standard evaluations that mix all capabilities risk measuring training data leakage or overlap specific to the benchmark rather than generalizable pure in-context learning capabilities. If the goal is to assess an LLM's ability to generalize to novel chemical spaces, benchmarks must isolate in-context learning capabilities. Our blinding protocol offers a concrete method for this: high performance that survives blinding (like Delaney) indicates robust reasoning, while performance that collapses (like Lipophilicity) indicates contamination or shallow heuristics.

For practical applications, the choice of which capabilities to leverage depends critically on the specific task requirements. General in-context learning is universally valuable, as it enables prediction on arbitrary structure−property relationships without domain-specific priors. Chemical in-context learning becomes particularly useful when the task involves molecular property prediction where the structural representation (e.g., SMILES) is familiar to the LLM. While not empirically validated in this work, we hypothesize this extends to other domains where the LLM has been exposed to the structural language during pretraining. Learned domain-specific relationships represent a double-edged sword: when the property is well-known to the LLM, leveraging this prior knowledge can substantially improve prediction performance. However, as demonstrated in our results, this same prior knowledge can diminish performance when coupled with small in-context data sets, particularly when the LLM's priors conflict with the provided examples. In such cases, strategically blinding the property name to suppress misleading priors may be beneficial.

One interesting finding is that the scale of the numbers to predict significantly affected the prediction performance at least in the case of QM7. This suggests that, in practice, rescaling the labels—as is routinely done for simple neural networks—might benefit the LLMs' predictions.

Limitations and Future Work

Our study has several limitations that should be considered when interpreting results. First, we use only SMILES string representations, while many molecular property prediction methods leverage 3D coordinates or molecular graphs. This choice reflects the text-based nature of LLMs but may underestimate their potential with richer input formats. Second, our statistical power is limited by computational costs: we perform only 2 runs per configuration. We quantify the resulting uncertainty by bootstrapping over the test set (Supporting Information) and have softened model-to-model comparisons accordingly; nonetheless, additional runs would tighten the confidence intervals and are a natural extension. Third, a study using an LLM with an openly available training corpus would make it possible to prove to what extent LLMs memorize specific numbers and data sets, which we cannot do here. Fourth, our main blinding sweep is confined to three widely used benchmarks (ESOL, Lipophilicity, QM7) that all predate 2018. This is a deliberate choice—these data sets are the most likely to exhibit contamination—but it raises the question of whether the conclusions transfer to newer data. As a first check, our 2025 postcutoff antiviral-potency control behaves just like the legacy sets (coincidence-level match and retention rates), suggesting the memorization conclusions are not an artifact of data set age. But, as the memorization of the positive controls shows, there are data sets from which the LLMs do recall values verbatim. The difference is most likely their appearance or nonappearance in the training corpus, which is a property of the data set−model pair rather than of either in isolation. On the mechanism behind this we can only speculate, since the training corpora are not accessible: what may decide it is not the age of a data set but how its values are published, with the boiling points and atomic weights tabulated as plain text across many independent pages whereas the benchmark values circulate as separate data files. Hence, the findings in this study need to be reevaluated for each new data set−model pair and indicate only that these LLMs have not been trained directly on these data sets. Regarding the other findings in our study, we expect the LLMs' in-context learning capabilities to improve with the general development of LLMs and hence the knowledge conflicts to diminish with time. But this is a hypothesis that needs to be tested in future work, for example by repeating the experiments presented in this study with future LLMs and data sets.

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

https://doi.org/10.1021/acs.jcim.6c00961
J. Chem. Inf. Model. 2026, 66, 12305−12315

CONCLUSION

In this work, we investigated whether large language models perform genuine in-context regression for molecular property prediction or whether their apparent success is driven by memorization or prior domain knowledge and in-context learning capabilities. Using a systematic six-level blinding framework across multiple data sets, and a memorization detector validated against a positive control of known values, we find no evidence that LLMs rely on direct, verbatim retrieval for the widely used molecular data sets predating 2018. Where prior knowledge contributes, our data are most consistent with generalized domain understanding rather than data set-specific recall—although a verbatim-match detector cannot exclude approximate recall or substructure-level statistical associations, so we state this as the reading our evidence supports rather than as a demonstrated mechanism. For values from other sources like Wikipedia and IUPAC, by contrast, LLMs do retrieve them verbatim.

Our results show that LLMs flexibly combine prior domain knowledge and in-context information, with this interaction both enabling and hindering performance depending on the task and the in-context sample count. In several cases, suppressing prior knowledge through blinding improves predictive accuracy, indicating that strong priors can interfere with in-context learning rather than support it. These effects vary substantially across properties and data sets, highlighting the limitations of standard benchmark evaluations. Often, giving the LLM access to 60 samples reduced the correlation compared to the 0-shot setting, while the addition of 1000 samples on average led to a significant improvement, at least for the frontier models.

Overall, our findings suggest that strong benchmark performance should not be equated with robust in-context regression. To test the pure in-context-learning abilities of LLMs, we recommend fully blinding the input data. For maximum predictive accuracy, however, whether including prior knowledge hinders performance—in the case of knowledge conflicts—or improves it depends on the specific case. Finally, rescaling the labels can further increase the predictive accuracy.

ASSOCIATED CONTENT

Data Availability Statement

All code, data, and precomputed results required to reproduce the experiments and figures in this paper are publicly available at https://github.com/MatthiasHBusch/BlindingLLMs.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.6c00961.

Figure S1: cumulative error distributions of the 0-shot predictions for Lipophilicity and QM7; Table S1: per-model memorization diagnostics on the three legacy benchmarks (3-significant-digit match rate, the two precision-scaling retention ratios m_2/m_1 and m_3/m_2, and Pearson r); Table S2: the same per-model diagnostics on the three control data sets (the 2025 postcutoff antiviral-potency negative control, and the IUPAC atomic-weight and Wikipedia boiling-point positive controls); Figure S2: distribution of the second significant digit of the QM7 versus Delaney ground-truth values, evidencing the value clustering behind QM7's elevated m_2/m_1 retention; Table S3: absolute three-significant-digit exact-match counts for all nine models across all six data sets; Table S4: the blinding levels applied to both positive controls—the 45 Wikipedia boiling points at all six levels and the 41 IUPAC atomic weights at the four levels defined for them—reporting correlation, verbatim-match rate, and retention per level (Section "Blinding the Positive Control"); Table S5: per-model precision-scaling retention with cluster-bootstrap confidence intervals for the boiling-point sweep, against the second-digit coincidence floor and beside each model's 0-shot retention on the same set; Table S6 and Figure S3: nonmonotonic (sine) transform control—affine-vs-sine correlations for GPT-5, GPT-4.1, and Gemini 2.5 Pro on Delaney at the three transformed blinding levels, against the kNN-Tanimoto structural and recall ceilings; Table S7: k-nearest-neighbor Tanimoto similarity-weighted baseline across neighbor counts; Table S8: trivial-feature (element- and character-count) ridge-regression baseline; Table S9: bootstrap 95% confidence intervals over the test set (Gemini 2.5 Pro); Table S10: statistical-significance (consistency-test) analysis of the main directional claims; Full prompt templates and design rationale for the 0-shot prompt and all six blinding levels (PDF)

AUTHOR INFORMATION

Corresponding Authors

Matthias Busch − *Institute for Continuum and Material Mechanics, Technical University of Hamburg, Hamburg 21073, Germany*; orcid.org/0009-0004-0420-119X; Email: matthias.busch@tuhh.de

Christian Feiler − *Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht 21502, Germany; Institute for Interface Physics and Engineering, Hamburg University of Technology, Hamburg 21073, Germany*; Email: christian.feiler@hereon.de

Roland C. Aydin − *Institute for Continuum and Material Mechanics, Technical University of Hamburg, Hamburg 21073, Germany; German Center for Artificial Intelligence (DFKI), Saarbrücken 66123, Germany; Saarland University, Saarbrücken 66123, Germany*; Email: roland.aydin@hereon.de

Authors

Marius Tacke − *Institute of Material Systems Modeling, Helmholtz-Zentrum Hereon, Geesthacht 21502, Germany*

Sviatlana V. Lamaka − *Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht 21502, Germany*

Mikhail L. Zheludkevich − *Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht 21502, Germany*

Christian J. Cyron − *Institute for Continuum and Material Mechanics, Technical University of Hamburg, Hamburg 21073, Germany; Institute of Material Systems Modeling, Helmholtz-Zentrum Hereon, Geesthacht 21502, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.6c00961

Author Contributions

M.B.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing—Original Draft, Writing—

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026

Review and Editing, Visualization. M.T.: Writing—Review and Editing. S.V.L.: Writing—Review and Editing. M.L.Z.: Writing—Review and Editing. C.J.C.: Writing—Review and Editing. C.F.: Supervision, Writing—Review and Editing. R.C.A.: Conceptualization, Supervision, Writing—Review and Editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF). This publication was funded via project 535656357 from the Deutsche Forschungsgemeinschaft (DFG): https://gepris.dfg.de/gepris/projekt/535656357 Claude Code (Anthropic) was used for text editing and revision of this manuscript. All authors reviewed and take full responsibility for the final content.

■ REFERENCES

(1) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.et al.Language models are few-shot learners*arXiv***2020**

(2) Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.et al.A survey on in-context learning*arXiv***2024**

(3) Zhou, Y.; Li, J.; Xiang, Y.; Yan, H.; Gui, L.; He, Y.The mystery of in-context learning: A comprehensive survey on interpretation and analysis*arXiv***2024**

(4) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X.What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks*arXiv***2023**

(5) Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, C. A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell* **2024**, *6*, 525−535.

(6) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell* **2024**, *6*, 161−169.

(7) Busch, M.; Tacke, M.; Lamaka, S. V.; Zheludkevich, M. L.; Linka, K.; Cyron, C. J.; Feiler, C.; Aydin, R. C. Large language models predicting the corrosion inhibition efficiency of magnesium dissolution modulators. *Corros. Sci.* **2025**, *255*, 113080.

(8) Joe, C. Y.; Song, K.; Chang, R. Evaluating In-Context Learning in Large Language Models for Molecular Property Regression. *J. Comput. Chem.* **2026**, *47*, No. e70308.

(9) Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(10) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(11) Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; Zhang, C.Quantifying memorization across neural language models*arXiv***2023**

(12) Cheng, Y.; Chang, Y.; Wu, Y.A survey on data contamination for large language models*arXiv***2025**

(13) Sainz, O.; Campos, J.; García-Ferrero, I.; Etxaniz, J.; de Lacalle, O. L.; Agirre, E.NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark*arXiv***2023**

(14) Bordt, S.; Srinivas, S.; Boreiko, V.; von Luxburg, U. How much can we forget about data contamination? **2024**; .

(15) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.

(16) Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.et al.ChemLLM: A chemical large language model*arXiv***2024**

(17) Akyürek, E.; Schuurmans, D.; Andreas, J.; Ma, T.; Zhou, D.What learning algorithm is in-context learning? investigations with linear models*arXiv***2022**

(18) Chen, S.; Chen, Y.; Li, Z.; Jiang, Y.; Wan, Z.; He, Y.; Ran, D.; Gu, T.; Li, H.; Xie, T.et al.Benchmarking large language models under data contamination: A survey from static to dynamic evaluation*Proceedings Of The 2025 Conference On Empirical Methods In Natural Language Processing*ACL202510091−10109

(19) Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P.Lost in the middle: How language models use long contexts*Transactions Of The Association For Computational Linguistics*ACL202412157−173

(20) Hemmati, R. *QM7−Dataset*, 2024. https://huggingface.co/datasets/HR-machine/QM7-Dataset, Accessed: 2026 January 14.

(21) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(22) Scheen, J.; MacDermott-Opeskin, H.; Chodera, J.; Griffen, E.; von Delft, F.; Aschenbrenner, J. C.; Fearon, D.; Balcomb, B.; Marples, P.; Tomlinson, C. W. E.; et al. *ASAP-Polaris-OpenADMET Blind Challenge Data*; zenodo 2025. DOI: .

(23) Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.et al.Chain-of-thought prompting elicits reasoning in large language models*arXiv.***2022**

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(25) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E.Neural message passing for quantum chemistry*arXiv***2017**

Downloaded from pubs.acs.org/jcisd8/article-pdf/66/19/12305/6854578/acs.jcim.6c00961.pdf by TU HAMBURG-HARBURG user on 05 November 2026