ACS MATERIALS LETTERS

www.acsmaterialsletters.org

LETTER

Polymer Solubility Prediction Using Large Language Models

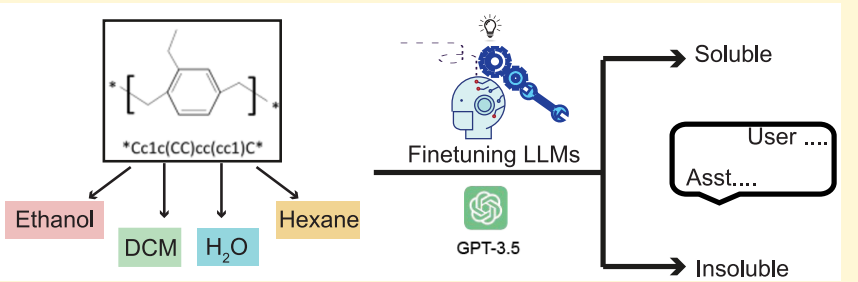
Published as part of ACS Materials Letters special issue "Machine Learning for Materials Chemistry".

Sakshi Agarwal, Akhlak Mahmood, and Rampi Ramprasad*

Cite This: *ACS Materials Lett.* 2025, 7, 2017−2023

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



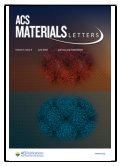
ABSTRACT: Traditional approaches in polymer informatics often require labor-intensive data curation, time-consuming preprocessing such as fingerprinting, and choosing suitable learning algorithms. Large language models (LLMs) represent a compelling alternative by addressing these limitations with their inherent flexibility, ease of use, and scalability. In this study, we propose a novel approach utilizing fine-tuned LLMs to classify solvents and nonsolvents for polymers, a property critical to polymer synthesis, purification, and diverse applications. Our results show that fine-tuned GPT-3.5 achieves predictive performance comparable to or exceeding traditional machine learning methods, even with limited data sets. The model achieved predictive accuracies of 0.90 and 0.83 for identifying soluble and insoluble solvent−polymer pairs, respectively. Remarkably, these models accurately classify solvents and nonsolvents in entirely unseen scenarios, indicating that they are able to effectively leverage the components embedded in their base models. The operational simplicity and accuracy of LLMs highlight their potential for advancing polymer research.

The solubility of polymers in various solvents is fundamental to the polymer industry, enabling diverse applications in coatings, adhesives, biomedical systems, and electronics.[1−6] Dissolved polymers exhibit essential properties for forming films, creating strong adhesive bonds, spinning fibers, or acting as encapsulating agents.[7] Conversely, polymers with low solubility find use in packaging and transport applications, exemplified by polyethylene's resistance to water, making it ideal for water storage and transport.[5,6]

Environmental and health considerations are increasingly guiding solvent selection.[7−9] The growing demand for eco-friendly, nontoxic solvents is driven by regulatory requirements and a societal focus on sustainability.[10] Water, which is a safe and environmentally benign solvent, is often preferred in biomedical applications for its excellent solubility properties and compatibility with sensitive biological systems. Thus, solvent selection is not only a technical necessity but also a key factor in achieving a sustainable and safe material design.

Despite the critical importance of solvent selection, determining the compatibility of a polymer with a specific solvent remains a challenging task. Over the past few years, many approaches leverage machine learning (ML) techniques to predict polymer−solvent interactions. These methods involve representing polymers and solvents using computational descriptors such as Polymer Genome, RDKit, or MACCS fingerprints.[11−19] While effective, these approaches often require significant effort in data preprocessing, model training, and interpretation, making them less accessible to experimentalists without specialized computational expertise. For instance, answering a straightforward question such as "Is polyethylene soluble in dichloromethane?" typically requires complex workflows involving polymer and solvent finger-

Received: January 8, 2025
Revised: April 21, 2025
Accepted: April 21, 2025
Published: April 25, 2025



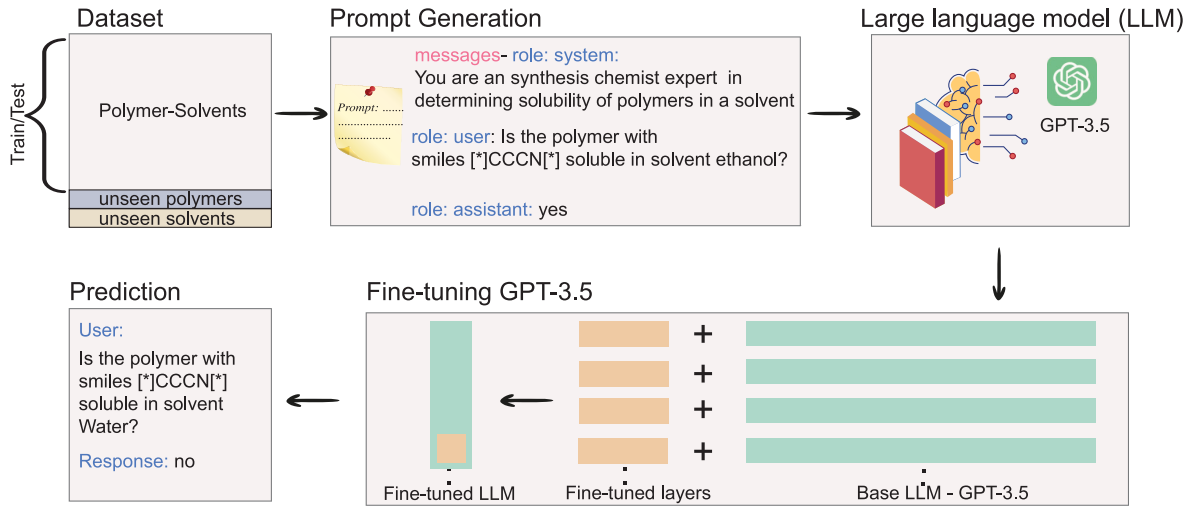


Figure 1. Fine-tuning workflow of GPT-3.5: The process begins with data set curation, proceeds to prompt generation, followed by fine-tuning the GPT-3.5 base model via the OpenAI API, and concludes in inference by the user.

printing, followed by the application of machine learning models. This labor-intensive process creates a bottleneck for users who could benefit from a simpler and more intuitive solution.

To address these challenges, we explored the potential of natural language processing (NLP) models, particularly large language models (LLMs). LLMs are especially appealing due to their ability to process natural language inputs and generate coherent, contextually relevant outputs.[20−24] This capability offers an intuitive interface, allowing people to pose questions in plain language. An intriguing question arises: can LLMs effectively address scientific queries, such as polymer−solvent compatibility, with the same ease as conversational prompts? Investigating this possibility opens the door to a transformative approach to predicting polymer solubility.

In this study, inspired by several recent efforts that leverage LLMs to solve materials problems,[24−28] we evaluated the capabilities of GPT-3.5, a state-of-the-art large language model developed by OpenAI,[29,30] for predicting polymer−solvent solubility. Using a data set comprising 3,373 polymers and 51 solvents, we fine-tuned GPT-3.5 to enable it to predict solubility with high accuracy. The fine-tuned model demonstrated good performance, accurately predicting the solubility for previously unseen polymers and solvents. This work represents a novel application of NLP models in materials science, showcasing their potential to simplify complex prediction tasks and improve accessibility for nonexperts. By eliminating the need for labor-intensive preprocessing and choice of learning algorithms, our approach bridges the gap between advanced computational tools and practical applications in polymer science. Furthermore, this study highlights the broader potential of LLMs in advancing data-driven research across chemical and material informatics.

The data set used in this fine-tuning study was previously obtained from a wide range of published sources, including peer-reviewed journals, printed handbooks, and online repositories.[18] The chemical space of the polymers included in the data set encompasses elements such as carbon (C), oxygen (O), selenium (Se), nitrogen (N), fluorine (F), phosphorus (P), sulfur (S), bromine (Br), silicon (Si), chlorine (Cl), iodine(I), boron (B), and hydrogen (H). To ensure clarity and consistency, the study excludes copolymers, polymer blends, polymers with additives, and cross-linked polymers. Additionally, this investigation focuses solely on room-temperature solubility, disregarding cases of partial solubility or solubility under high-temperature conditions. The data set comprises 6,282 polymers and 58 solvents, resulting in 19,649 soluble pairs and 9972 insoluble pairs, making a total of 30034 polymer solvent pairs. The data set does not provide exhaustive combinations of solvent−polymer pairs. For example, a given polymer might have solubility data with only five solvents, while another polymer might have data with entirely different solvents, as shown in Figure 1. In addition, some polymers and solvents had data exclusively for either solubility or insolubility, without any instances of both. The complete list of 58 solvents used in the study is provided in Table S1. Additionally, the polymers used in this study include a few common polymers, listed in Table S2. However, the majority of polymers lack specific names, so the SMILES representations of the polymers were used instead. This curated data set was then utilized to fine-tune the LLM, as discussed in the next subsection.

Large language models (LLMs) have demonstrated exceptional capabilities in handling tasks involving simple language prompts. This study aims to extend their applicability to a scientific challenge—classifying polymer−solvent solubility. To adapt these models to such specialized tasks, fine-tuning was employed. Fine-tuning involves training the base LLM model on a curated data set consisting of prompts and their corresponding responses, tailored to the specific problem domain. For this study, a data set comprising 30,034 polymer−solvent pairs was converted into question-answer format. The fine-tuning of GPT follows a structured format, as illustrated in Figure 1, involving three distinct roles in the input message. First is the system role, which defines the model's expertise and sets the context for the task. Second is the user role, where the input is framed as a question. Each data point in this study was converted into the question: "Is the polymer with 'SMILES' soluble in the solvent 'solvent name'?" Here, the placeholder 'SMILES' was replaced with the polymer's Simplified Molecular Input Line Entry System (SMILES) representation, and 'solvent name' was replaced with the corresponding solvent's name from the data set. Finally, the assistant role provides the correct answer to the questions for the training

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023

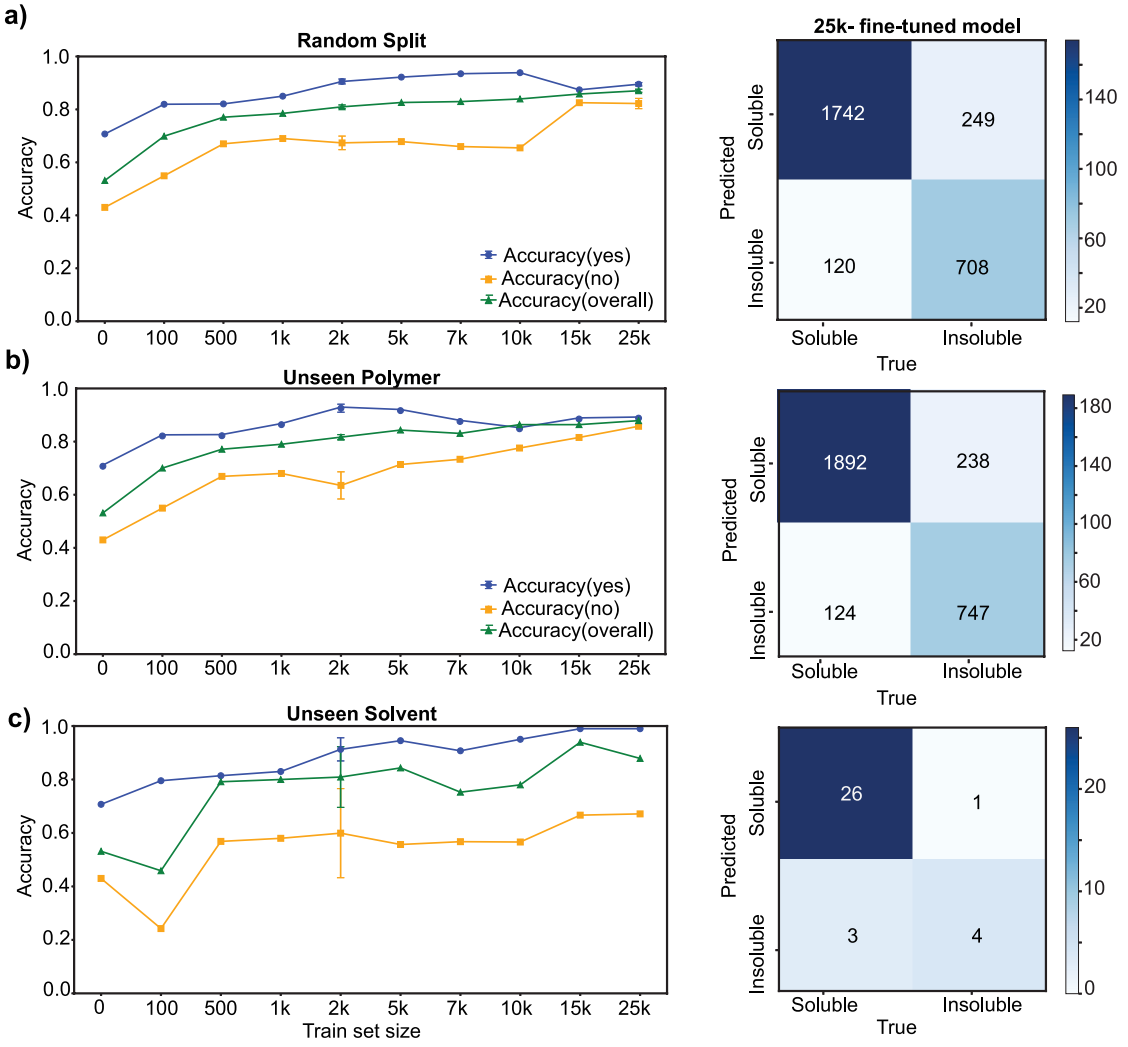


Figure 2. Class-wise accuracy for each finetuned model from 2000 question/answers to 25000 (a) for test data set having seen solvents, seen polymers, unseen polymers, and unseen solvents, (b) test set with only unseen polymers, and (c) test set with only unseen solvents. The confusion matrices for each test set correspond to the 25k-finetuned model.

set. The model was trained to respond with a text completion from a predefined set of possible responses: "yes" or "no", depicting soluble and insoluble polymers in the given solvent. This structured approach enabled the model to learn and respond accurately to solubility classification tasks during fine-tuning.

To evaluate the model's performance under varying training data set sizes, we fine-tuned the LLM on data sets of different sizes, ranging from 2,000 to 25,000 data points. Each subset was balanced to include an equal number of soluble and insoluble polymer−solvent pairs for training. The results of these experiments provide insights into the model's robustness and efficiency in low-data regimes.

A critical aspect of this study was to assess the model's ability to generalize to unseen polymers and solvents. The complete data set was partitioned such that 10% of the polymers were set aside as unseen polymers, ensuring they were entirely absent from the training set. Additionally, eight solvents were excluded from the training data and designated as unseen solvents. After isolating these two sets, the remaining data was divided into training and testing subsets. The training data sets were constructed incrementally (e.g., 0, 100, 500, 1000, 2,000, 5,000, 10,000, 15,000, and 25,000 rows), as discussed above. The remaining data and the unseen held out data are considered as test data. As a result, the test data contained examples of four distinct categories: seen polymers with seen solvents, seen polymers with unseen solvents, unseen polymers with seen solvents, and unseen polymers with unseen solvents. This systematic approach allowed for a comprehensive assessment of the model's ability to handle both familiar and novel scenarios. The framework for data set preparation and the systematic partitioning scheme for training and testing are visualized in Figure S1 and Figure 1, providing a clear outline of the methodology employed in this study.

Fine-tuning using the polymer−solvent database was conducted with GPT-3.5. This process involved training the base GPT-3.5 model on a task-specific data set. By fine-tuning of the model for the specific task, the weights of the layers were updated, resulting in a specialized and optimized language model, as illustrated in Figure 1. The procedure was both efficient and scalable, completing in just a few minutes for smaller data sets. Even with larger data sets, fine-tuning never exceeded 1 h, highlighting GPT-3.5's rapid adaptability. This capability demonstrates its practical utility for scientific tasks, enabling swift and effective customization.

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023

Once fine-tuned, the model was deployed for solubility classification tasks, including predictions on the reserved test set containing seen solvents, seen polymers, unseen polymers, and unseen solvents. This inference stage involved querying the fine-tuned model with test polymer−solvent pairs to classify their solubility. Similar to the training set, the test set was also formatted into the GPT message structure. However, the assistant role, which provides the response (Figure 1), was excluded to allow the model to generate predictions independently during testing. The process took a few minutes to hours to complete depending on the test set size. The results provided critical insights into the model's performance across various test scenarios, including its accuracy with seen and unseen data. Each fine-tuned model was evaluated on the test set using accuracy metrics designed to assess its performance in predicting polymer−solvent solubility. Two distinct accuracy scores were computed to separately measure the model's ability to classify soluble and insoluble cases. Accuracy (yes) evaluates the model's performance in correctly identifying soluble polymer−solvent pairs. It is calculated as the ratio of true positive cases (correctly predicted "yes" responses) to the total number of actual soluble cases in the data set:

$$\text{Accuracy (yes)} = \frac{\text{Correctly predicted soluble}}{\text{Total number of soluble}} \quad (1)$$

Accuracy (no) assesses the model's ability to correctly identify insoluble polymer−solvent pairs. It is computed as the ratio of true negative cases (correctly predicted "no" responses) to the total number of actual insoluble cases in the data set:

$$\text{Accuracy (no)} = \frac{\text{Correctly predicted insoluble}}{\text{Total number of insoluble}} \quad (2)$$

Both accuracy (yes) and accuracy (no) was calculated for the responses generated by base GPT model and the fine-tuned models. Figure 2 illustrates the accuracy for both yes (soluble) and no (insoluble) classifications with increasing data set size. It is evident from the figure that the accuracy is lowest for the base GPT model with a 0 train set size. The accuracy increases with an increasing number of training question-answer pairs. The results indicate a clear trend: the accuracy of the model improves consistently with an increase in the number of training examples across all test cases. For the test set, which includes combinations of seen and unseen polymers and solvents (Random split) as shown in Figure 2 (a), the model trained on 25,000 question-answer pairs has the highest accuracy for both classes. An accuracy (yes) of 0.9 and an accuracy (no) of 0.83 was achieved. A similar trend was observed when the model was tested specifically on unseen polymers (Figure 2 (b)): accuracy (yes) of 0.89 and accuracy (no) of 0.85, confirming the model's ability to generalize to completely novel polymers. This finding highlights the model's potential applicability to real-world scenarios where previously unencountered polymers require solubility predictions. In the case of unseen solvents, the model also achieved competitive accuracy with an accuracy (yes) of 0.99 and an accuracy (no) of 0.68 (Figure 2 (c)). This demonstrates the robustness of the fine-tuned language model in extrapolating its understanding to new solvent systems, even when those solvents were absent from the training data. Note, for base GPT model (0 train-set size), all the polymers and solvent are unseen; therfore, the accuracy values were kept same in all the three plots. The large error bars at a training size of 2k in the plot indicate high variability in model performance across different training subsets. This suggests that at this specific training size, the model may be particularly sensitive to the composition of the training data. Possible reasons include an imbalanced representation of solvents, leading to inconsistent learning outcomes, or the inherent difficulty of generalizing to unseen solvents with limited data of 2k. Additionally, this fluctuation could be due to instability in model convergence at intermediate training sizes. This further suggests that increasing data size enhances the performance of the model. To further assess the model's classification performance, confusion matrices for the best performing model (Figure 2) and other models (Figure S2 and S3) were generated. These matrices provide a detailed breakdown of the model's ability to classify soluble and insoluble polymer−solvent pairs. The high diagonal values in the matrices (representing correct classifications) affirm that the model effectively distinguishes between soluble and insoluble cases for all three train cases.

Our model performs comparable to existing machine learning models.[11−18] As shown in Figure S4, for a random split, the accuracies for "yes" (0.90) and "no" (0.83) predictions closely match the F1 scores reported by Kern et al.[18] using random forest classifier (RF) (0.92 and 0.89, respectively). Similar performance trends are observed for unseen polymers. However, a notable improvement is evident in the case of the unseen solvents. After fine-tuning on 25,000 training question-answer pairs, our model achieves an accuracy of 0.99 for predicting solubility and 0.69 for predicting insolubility, exceeding the previously reported F1 scores of 0.73 and 0.67, respectively. While this analysis highlights that the fine-tuned model performs on par with conventional machine learning models and demonstrates superior performance on unseen data, it is important to note that a direct comparison is limited due to differences in the training data and data-splitting methods.

The fine-tuning process involved optimizing the number of training epochs and the softmax temperature during inference to maximize the model's performance. Figure S5 presents the effect of varying the number of epochs, starting from 5 epochs, on the model's accuracy. Training the model for different epochs revealed that 10 epochs provided the most optimal balance between performance and computational efficiency for the data set used in this study. While increasing the number of epochs beyond 10 did not significantly improve accuracy, it introduced the risk of overfitting. In addition to epochs, the softmax temperature, a key parameter in text generation, was tuned to evaluate its impact on prediction accuracy. The softmax temperature controls the imaginativeness of the model's text generation. A lower value (e.g., 0.1) makes the model more deterministic, focusing on high-probability outputs. On the other hand, a higher value (e.g., 1.0) increases randomness, allowing the model to explore diverse outputs but at the risk of generating less accurate or coherent predictions. For this study, a temperature of 0.5 was found to be the most efficient as shown in Figure S5, yielding the highest accuracy across test scenarios. At this setting, the model maintained a balance between deterministic and probabilistic behavior, ensuring consistent and reliable predictions.

To gain deeper insights into the best model's performance for each solvent and its ability to differentiate between soluble and insoluble cases, accuracy was calculated using the best performing model (25k-model) individually for each solvent in the data set as shown in Figure 3. The figure provides a

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023

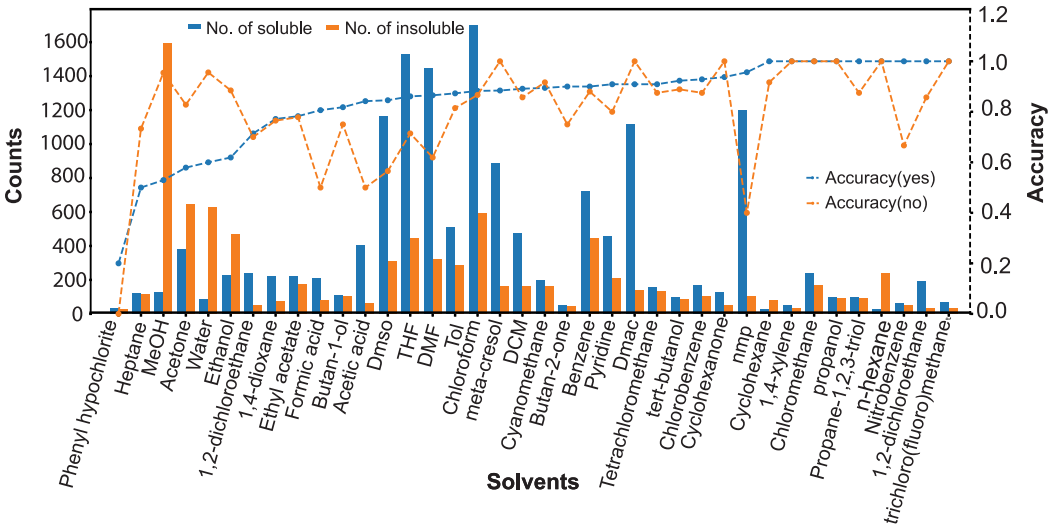


Figure 3. Accuracy (yes) and Accuracy (no) for all seen solvents, correlated with the data distribution of each solvent in the training set. Blue and orange bars depict the number of data points corresponding to that solvent in the training set. The blue and orange circles represent the corresponding accuracy (yes) and accuracy (no), respectively.

comprehensive view of the relationship between data distribution and model accuracy. The bars in Figure 3 represent the number of data points per solvent in the training data set, categorized as soluble (blue) and insoluble (orange). The corresponding accuracy scores for each class (accuracy (yes) for soluble cases and accuracy (no) for insoluble cases) are depicted as circles. Solvents with a higher number of data points for a particular class tend to show a higher accuracy for that class in most cases. For instance, methanol (MeOH), water, and acetone have a larger proportion of insoluble cases in the training data set, resulting in higher accuracy (no). Conversely, chloroform, dimethylformamide (DMF), and tetrahydrofuran (THF) have more soluble cases, leading to an accuracy (yes). Looking closely at the accuracy and the number of data points per each solvent, it is evident that most of the protic solvents have lower accuracy (yes) compared to accuracy (no). Therefore, the solvents were divided into protic and aprotic solvents, and the accuracies were calculated. It is clear from the Figure 4(a) that the aprotic solvents had a larger number of soluble cases in the training data, resulting in higher accuracy (yes) for the test set. On the other hand, protic solvents had more insoluble cases, which translated to higher accuracy (no). This analysis highlights the influence of the distribution of solubility data on the resulting model performance. Similar trends were observed across the other fine-tuned models trained with varying amounts of data, ranging from 2,000 to 15,000 train set size as shown in supplementary Figure S6 and S7. In each case, the model's performance consistently improved with an increasing volume of training data for each class per solvent type, mirroring the patterns seen in for 25,000 questions. This highlights a clear correlation between the distribution of the training data set and the model's accuracy in predicting solubility.

Furthermore, for unseen solvents, the results (Figure 4(b)) show a generally high accuracy (yes), demonstrating the model's capability to generalize to new solvents. The model is performing well for a few of the unseen solvents, which appears to be similar to the solvents present in the training data set. This suggests that the model is able to establish a correlation between similar types of solvents. However, a few cases exhibit slightly lower accuracy, primarily due to an imbalance in the training data. For majority of the solvents in the training data there are fewer insoluble cases, evident from Figure 3, where orange bars are smaller than blue bars. This imbalance affects the model's performance, leading to slightly reduced accuracy (no) for these unseen solvents. One notable exception is the unseen solvent n-[bis(dimethylamino)phosphoryl]-n-methyl-methanamine (nbm), which is entirely different in chemical nature from all of the seen solvents present in the training data. The model struggles to establish correlations between nbm and the seen solvents, resulting in the poorest performance for this solvent. This indicates that the model's extrapolation ability is limited when faced with chemically distinct solvents that deviate significantly from the training examples. This analysis underscores the model's ability to understand and generalize solubility relationships from the SMILES representation of polymers and solvents. The strong performance across seen and unseen cases confirms the model's utility in practical applications, while the observed limitations highlight areas for improvement, particularly in handling underrepresented or chemically unique solvents.

The fine-tuning of the GPT model comes with associated costs, as charged by OpenAI. These costs depend on the number of tokens or rows used during the fine-tuning and inference stages. The cost of fine-tuning is $0.0036 per row. For example, fine-tuning the model on 25,000 rows of polymer−solvent data (including corresponding responses) amounts to $90. This cost scales linearly with the number of rows, meaning smaller data sets incur proportionally lower costs. Inference costs are significantly lower than fine-tuning costs, at $0.0005 per row. For instance, processing 5,000 rows during the inference stage costs $2.50. In total, we fine-tuned six different models, each trained on an increasing number of rows, ranging from 2,000 to 25,000. The cumulative cost for fine-tuning and inference across all models amounted to approximately $400. While the use of GPT models demonstrates promising results in solving scientific problems like solubility prediction, the associated costs may be a limiting factor, particularly for large-scale data sets or iterative fine-tuning efforts. However, this study also opens doors for

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023

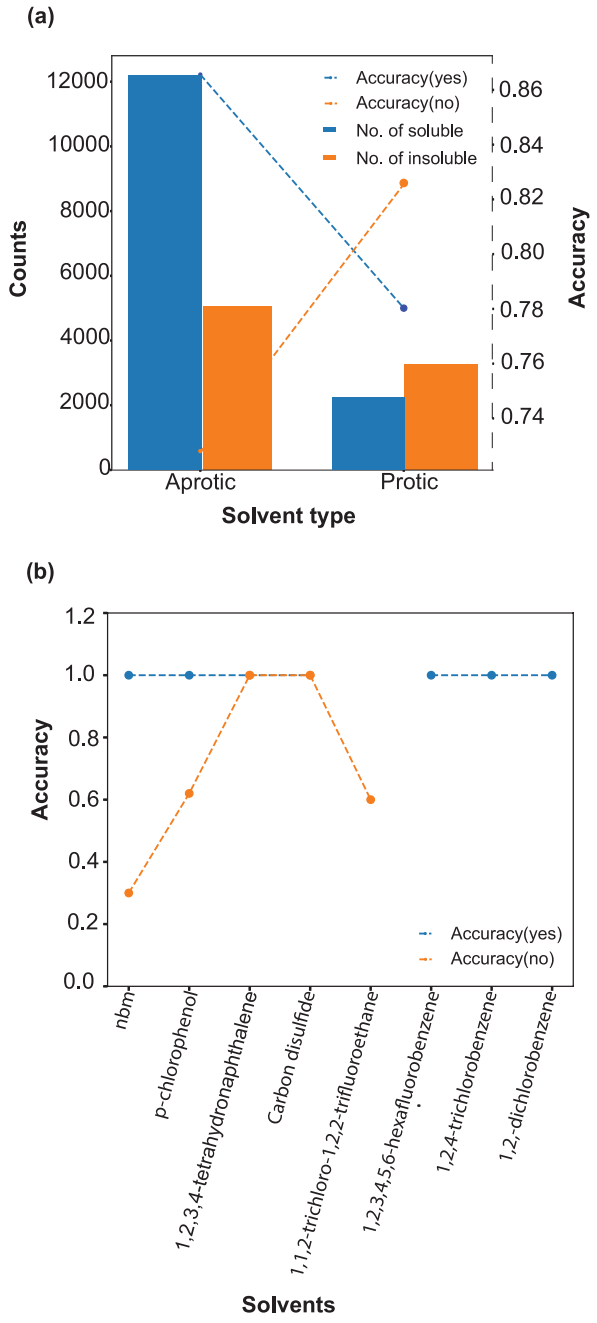


Figure 4. (a) Class-wise accuracy for protic and aprotic solvents based on their data distribution in the training set. Blue and orange bars depict the number of data points corresponding to that solvent in the training set. The blue and orange circles represent the corresponding accuracy (yes) and accuracy (no), respectively. (b) Accuracy for each unseen solvent.

exploring open-source alternatives such as LLaMA, Falcon, or other publicly available large language models. These models offer the potential to significantly reduce costs while maintaining competitive performance levels, especially when fine-tuned and deployed on a local or cloud-based infrastructure.

In terms of runtime, the Random Forest (RF) model including hyperparameter tuning took approximately 9 min, while polymer genome fingerprinting for 80% of the data required an additional 30 min. In comparison, fine-tuning GPT on 80% of the data took around 45 min. The times required for both approaches are thus comparable. However, GPT offers the added advantage of a more user-friendly implementation, making it accessible for broader applications.

This study explores the use of fine-tuned large language models (LLMs), specifically GPT-3.5, for predicting polymer solubility in solvents, a fundamental property in polymer science. Starting with a curated data set of 30,034 polymer−solvent pairs, the data was reformulated into question-answer prompts to enable model fine-tuning. To systematically evaluate model performance on both seen and unseen polymer−solvent combinations, fine-tuning was conducted on varying data set sizes. The results demonstrated increasing accuracy with larger training data sets, with fine-tuned models achieving comparable or superior performance to traditional methods without requiring complex preprocessing like fingerprinting and choosing learning algorithms. Performance analysis revealed that the solvent type and data distribution were critical factors influencing accuracy, and the model exhibited strong generalization capabilities for previously unseen cases. These findings underscore the ability of fine-tuned LLMs to directly infer solubility relationships from SMILES representations of polymers and solvents, offering significant practical utility. The demonstrated feasibility and efficiency of LLMs in material informatics position them as a promising alternative to conventional machine learning methods.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsmaterialslett.5c00054.

Table containing detailed description of all the solvents present and their corresponding data points. Schematic for train-test split strategy. Confusion matrices corresponding to all the fine-tuned models. Class-wise accuracy with respect to number of epochs and softmax temperature of inference. Solvent-wise accuracy and the corresponding data distribution (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Rampi Ramprasad − *School of Materials Science and Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, Georgia 30318, United States;* orcid.org/0000-0003-4630-1565; Email: rampi.ramprasad@mse.gatech.edu

Authors

Sakshi Agarwal − *School of Materials Science and Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, Georgia 30318, United States*

Akhlak Mahmood − *School of Materials Science and Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, Georgia 30318, United States;* orcid.org/0000-0002-5607-2885

Complete contact information is available at: https://pubs.acs.org/10.1021/acsmaterialslett.5c00054

Author Contributions

CRediT: **Sakshi Agarwal** conceptualization, formal analysis, investigation, methodology, writing - original draft; **Akhlak U.**

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023

Mahmood conceptualization, methodology, writing - review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research through grants N00014-19-1-2103 and N00014-20-1-2175.

■ REFERENCES

(1) Miller-Chou, B. A.; Koenig, J. L. A review of polymer dissolution. *Prog. Polym. Sci.* **2003**, *28*, 1223−1270.
(2) Acikgoz, C.; Hempenius, M. A.; Huskens, J.; Vancso, G. J. Polymers in conventional and alternative lithography for the fabrication of nanostructures. *Eur. Polym. J.* **2011**, *47*, 2033−2052.
(3) Dong, X.; Lu, D.; Harris, T. A.; Escobar, I. C. Polymers and solvents used in membrane fabrication: a review focusing on sustainable membrane development. *Membranes* **2021**, *11*, 309.
(4) Choi, M.-J.; Woo, M. R.; Choi, H.-G.; Jin, S. G. Effects of polymers on the drug solubility and dissolution enhancement of poorly water-soluble rivaroxaban. *International Journal of Molecular Sciences* **2022**, *23*, 9491.
(5) Medarević, D.; Djuriš, J.; Barmpalexis, P.; Kachrimanis, K.; Ibrić, S. Analytical and computational methods for the estimation of drug-polymer solubility and miscibility in solid dispersions development. *Pharmaceutics* **2019**, *11*, 372.
(6) Kadajji, V. G.; Betageri, G. V. Water soluble polymers for pharmaceutical applications. *Polymers* **2011**, *3*, 1972−2009.
(7) Tsampanakis, I.; Orbaek White, A. The mechanics of forming ideal polymer−solvent combinations for open-loop chemical recycling of solvents and plastics. *Polymers* **2022**, *14*, 112.
(8) Rivas, B. L.; Urbano, B. F.; Sánchez, J. Water-soluble and insoluble polymers, nanoparticles, nanocomposites and hybrids with ability to remove hazardous inorganic pollutants in water. *Frontiers in chemistry* **2018**, *6*, 320.
(9) Duis, K.; Junker, T.; Coors, A. Environmental fate and effects of water-soluble synthetic organic polymers used in cosmetic products. *Environmental Sciences Europe* **2021**, *33*, 21.
(10) Barton, A. F. Solubility parameters. *Chem. Rev.* **1975**, *75*, 731−753.
(11) Thakral, S.; Thakral, N. K. Prediction of drug−polymer miscibility through the use of solubility parameter based Flory−Huggins interaction parameter and the experimental validation: PEG as model polymer. *Journal of pharmaceutical sciences* **2013**, *102*, 2254−2263.
(12) Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian approach to predict solubility parameters. *Advanced Theory and Simulations* **2019**, *2*, 1800069.
(13) Kurotani, A.; Kakiuchi, T.; Kikuchi, J. Solubility prediction from molecular properties and analytical data using an in-phase deep neural network (Ip-DNN). *ACS omega* **2021**, *6*, 14278−14287.
(14) Chi, M.; Gargouri, R.; Schrader, T.; Damak, K.; Maâlej, R.; Sierka, M. Atomistic descriptors for machine learning models of solubility parameters for small molecules and polymers. *Polymers* **2022**, *14*, 26.
(15) Liu, T.-L.; Liu, L.-Y.; Ding, F.; Li, Y.-Q. A machine learning study of polymer-solvent interactions. *Chin. J. Polym. Sci.* **2022**, *40*, 834−842.
(16) Chandrasekaran, A.; Kim, C.; Venkatram, S.; Ramprasad, R. A deep learning solvent-selection paradigm powered by a massive solvent/nonsolvent database for polymers. *Macromolecules* **2020**, *53*, 4764−4769.
(17) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports* **2021**, *144*, 100595.
(18) Kern, J.; Venkatram, S.; Banerjee, M.; Brettmann, B.; Ramprasad, R. Solvent selection for polymers enabled by generalized chemical fingerprinting and machine learning. *Phys. Chem. Chem. Phys.* **2022**, *24*, 26547−26555.
(19) Tran, H.; Gurnani, R.; Kim, C.; Pilania, G.; Kwon, H.-K.; Lively, R. P.; Ramprasad, R. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials* **2024**, *9*, 866.
(20) Zhang, X.; Zhou, Z.; Ming, C.; Sun, Y.-Y. GPT-Assisted Learning of Structure−Property Relationships by Graph Neural Networks: Application to Rare-Earth-Doped Phosphors. *J. Phys. Chem. Lett.* **2023**, *14*, 11342−11349.
(21) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 all you need for low-data discovery in chemistry? *ChemRxiv* 2023. https://chemrxiv.org/engage/chemrxiv/article-details/63eb5a669da0bc6b33e97a35.
(22) Xie, T.; Wan, Y.; Huang, W.; Zhou, Y.; Liu, Y.; Linghu, Q.; Wang, S.; Kit, C.; Grazian, C.; Zhang, W. Large language models as master key: unlocking the secrets of materials science with GPT. *arXiv* 2023, arXiv:2304.02213. https://arxiv.org/abs/2304.02213.
(23) Dai, D.; Zhang, G.; Wei, X.; Lin, Y.; Dai, M.; Peng, J.; Song, N.; Tang, Z.; Li, S.; Liu, J.; et al. A GPT-assisted iterative method for extracting domain knowledge from a large volume of literature of electromagnetic wave absorbing materials with limited manually annotated data. *Comput. Mater. Sci.* **2025**, *246*, 113431.
(24) Xie, Z.; Evangelopoulos, X.; Omar, Ö. H.; Troisi, A.; Cooper, A. I.; Chen, L. Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chemical science* **2024**, *15*, 500−510.
(25) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **2024**, *6*, 161−169.
(26) Mishra, V.; Singh, S.; Ahlawat, D.; Zaki, M.; Bihani, V.; Grover, H. S.; Mishra, B.; Miret, S.; Krishnan, N. Foundational Large Language Models for Materials Research. *arXiv* 2024, arXiv:2412.09560. https://arxiv.org/abs/2412.09560.
(27) Ock, J.; Guntuboina, C.; Barati Farimani, A. Catalyst energy prediction with catberta: Unveiling feature exploration strategies through large language models. *ACS Catal.* **2023**, *13*, 16032−16044.
(28) Ramos, M. C.; Collison, C.; White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science* **2025**, *16*, 2514−2572.
(29) Lee, J.-S.; Hsiang, J. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information* **2020**, *62*, 101983.
(30) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877−1901.

https://doi.org/10.1021/acsmaterialslett.5c00054
ACS Materials Lett. 2025, 7, 2017−2023