Supporting Information

Prompt2Poly: Ask, Specify, Create – A Dialogue-Based Large Language Model for Targeted TSMPs Design

Borun Das,*,† Andrew Peters,‡ Guoqiang Li,¶ and Xiali Hei*,†

†*Department of Computer Science, University of Louisiana at Lafayette, Louisiana, USA*

‡*Department of Chemical Engineering, Louisiana Tech University, Ruston, Louisiana, USA*

¶*Department of Mechanical Engineering, Louisiana State University, Baton Rouge, Louisiana, USA*

E-mail: borun.das1@louisiana.edu; xiali.hei@louisiana.edu

1 Property Prediction Models

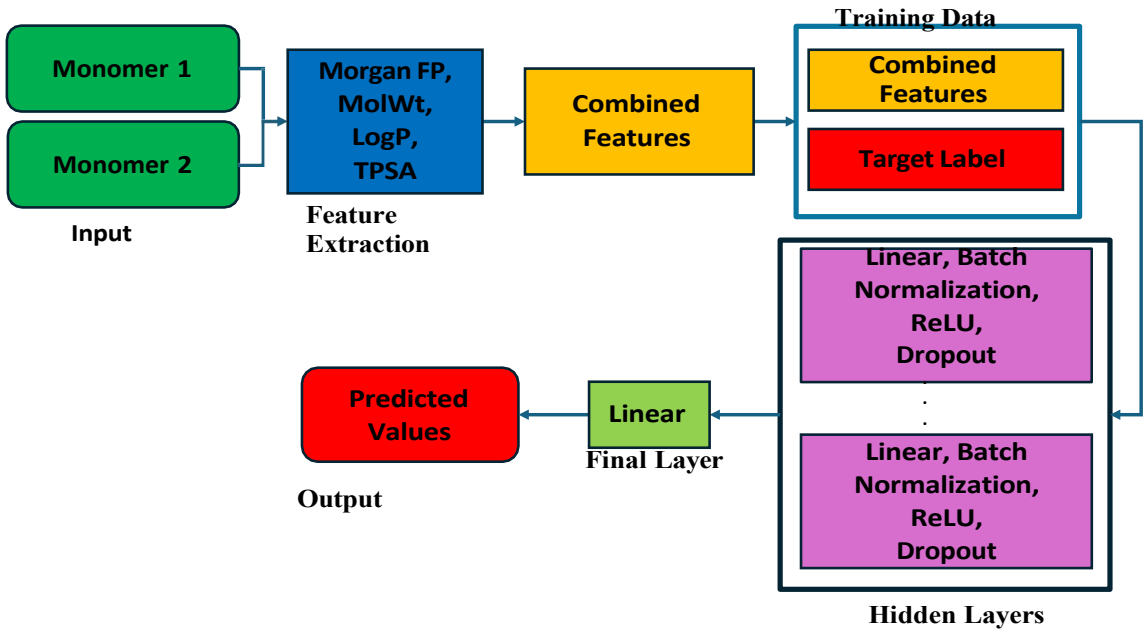

Figure S1: Flow Diagram of the Property Prediction Models

We built two separate multilayer perceptron (MLP) models to predict glass transition temperature (T_g) and rubbery modulus (E_r). Both models share the same architecture but are trained on different targets. Each network is implemented as a sequence of three hidden layers, where each layer consists of a Linear transformation, Batch Normalization, ReLU activation, and Dropout. The hidden layer sizes are 256, 128, and 64 units, with dropout rates of 0.5, 0.4, and 0.3, respectively. Kaiming initialization is applied to all linear layers to improve training stability. The flow diagram of these models is illustrated in Figure S1. We extract input features from two monomers. For each monomer, we compute Morgan fingerprints (radius 2, size 200) and molecular descriptors using RDKit. The descriptors include molecular weight (MolWt), LogP, and topological polar surface area (TPSA). We combine the fingerprints and descriptors of both monomers, and we also include the mixing ratios. This gives us a 208-dimensional feature vector. We scale all features using StandardScaler. We use the mean squared error (MSE) as the loss function. Each model is trained separately using the AdamW optimizer (learning rate = 0.001). We apply

a ReduceLROnPlateau scheduler to lower the learning rate when validation loss stops improving. We split the dataset into 80% training and 20% validation. We save the model weights that give the best validation loss. The training and validation losses for each model are shown in Figure S2 (T_g) and Figure S3(E_r). Both models show low loss, which means they can learn the property patterns from the data.

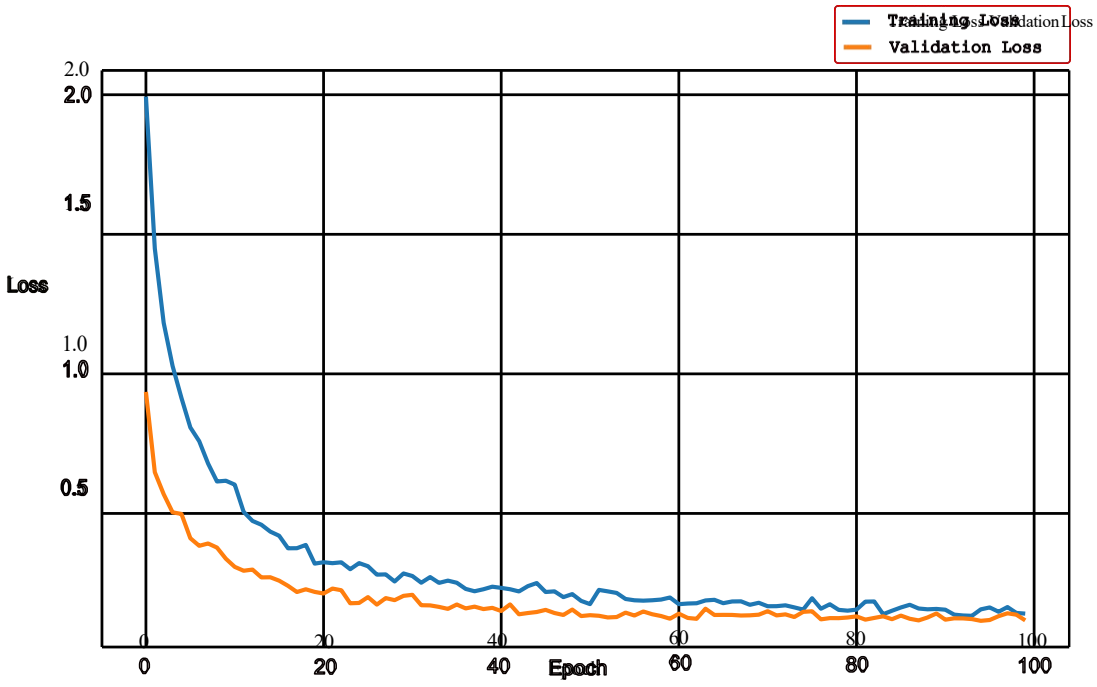

Figure S2: Training and validation loss for T_g prediction model.

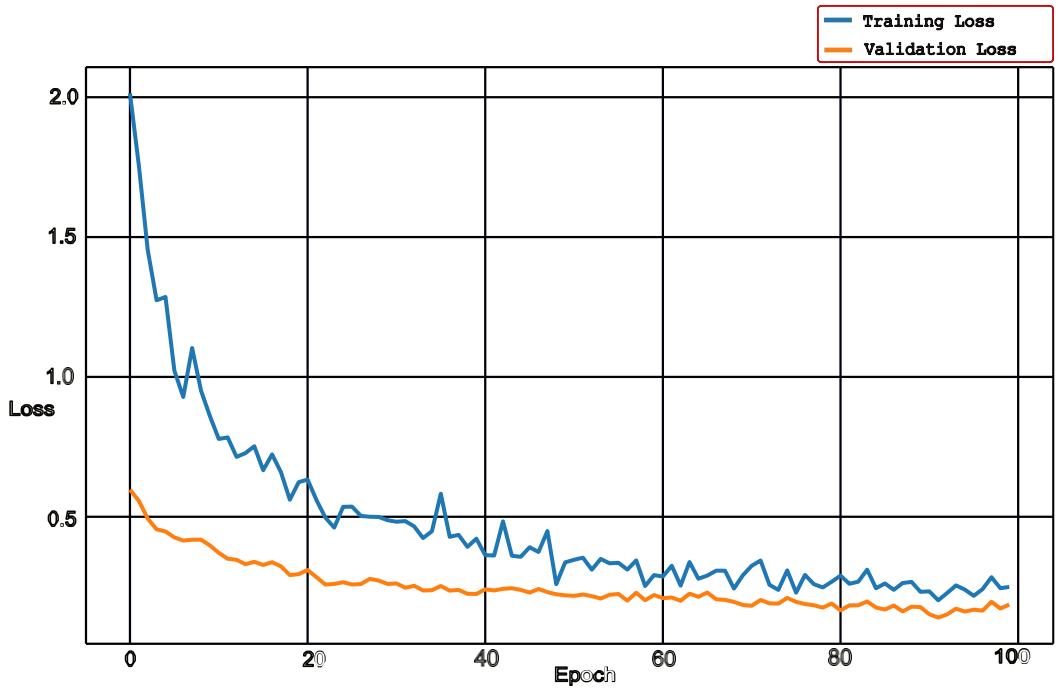

Figure S3: Training and validation loss for E_r prediction model.

2 Additional Case Studies in Conversational TSMPs Design

To further demonstrate the capabilities of our framework, we present additional user–model interactions for generating TSMPs. These examples show how the fine-tuned language model interprets natural language prompts that specify target properties or functional groups and returns structurally valid polymer candidates. Each example includes the user prompt, the model's response, and the corresponding SMILES representations, with candidate polymers provided in each case. The examples are not filtered for uniqueness, and some outputs may therefore appear as duplicates. The primary objective is to illustrate the framework's ability to capture user intent and produce chemically reasonable polymer structures in a conversational setting.

Examples S1-S3 present selected multi-turn dialogues, along with the corresponding polymer candidates generated by the GPT-4o-mini fine-tuned model.

Example S1

User: I want to make a thermoset shape memory polymer

Assistant: Would you like to design your TSMP based on certain functional groups, or are you more interested in achieving specific chemical properties?

User: both

Assistant: I can help with various functional groups (like epoxy, imine, vinyl, thiol) and specific properties (Tg and Er). What's your priority - groups or properties?

User: Please give me some TSMP with Tg = 100C and Er= 40MPa and vinyl(C=C) groups in monomer 1 and vinyl(C=C) groups in monomer 2

Assistant: Two TSMPs for you (as I told model to generate two samples per query) Sample : 1 Here's a valid monomer pair designed with vinyl groups for crosslinking and optimized for your Tg and Er targets:

Monomer 1: CC(C)(c2ccc(OCC1CO1)cc2)c4ccc(OCC3CO3)cc4

Monomer 2: CC(N)COCC(C)OCC(C)OCC(C)OCC(C)N

Molecular structures of these monomers are given below:

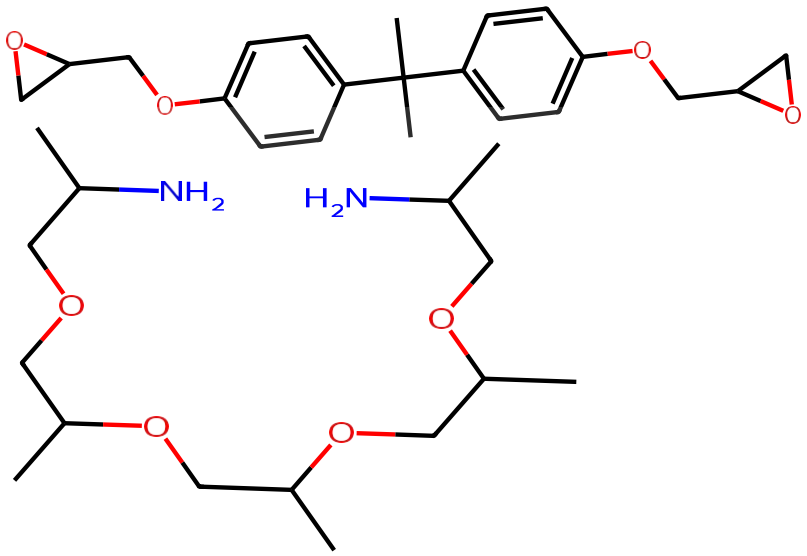

Figure S4: Molecular structures of Example 1

Example S2

User: Please suggest me some TSMPs

Assistant: What matters most for your TSMP design - the functional groups we use or the target properties?

User: Please focus on group based monomer pairs

Assistant: I can create TSMPs using functional groups such as epoxy (C1OC1), imine (NC), vinyl (C=C), or thiol (CCS), while meeting specific Tg/Er requirements.

User: Please give me some TSMP with epoxy(C1OC1) groups in monomer 1 and imine (NC) groups in monomer 2

Assistant: Two TSMPs for you (as I told model to generate two samples per query) Sample: 1 Here's a pair of monomers fulfilling the criteria of epoxy groups and imine groups:

Monomer 1: Cc1cc(OCCOCC2CO2)ccc1OCCOCC4CO4

Monomer 2: CC1(C)CC(N)CC(C)(CN)C1

Molecular structures of these monomers are given below:

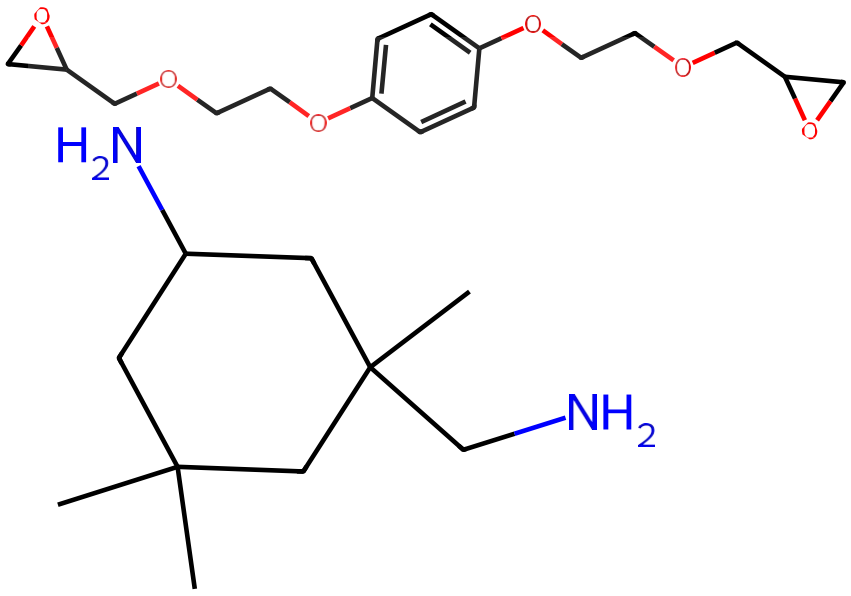

Figure S5: Molecular structures of Example 2

Example S3

User: I want to make a thermoset shape memory polymer

Assistant: Do you have particular functional groups in mind, or are you looking for a certain property set?

User: Please focus on property based monomer pairs

Assistant: I'll need your target values for Tg (glass transition temp) and Er (rubbery modulus) to proceed.

User: Please give me some TSMP with Tg = 100C and Er= 100MPa

Assistant: Two TSMPs for you (as I told model to generate two samples per query) Sample: 1

Both monomers I suggested are compatible and can be easily accessed: Monomer 1: CC(C)(c2ccc(OCC1CO1)cc2)c4ccc(OCC3CO3)cc4

Monomer 2: CCCC(=O)OCC(CC)(COC(=O)CCS)COC(=O)CCS

Molecular structures of these monomers are given below:

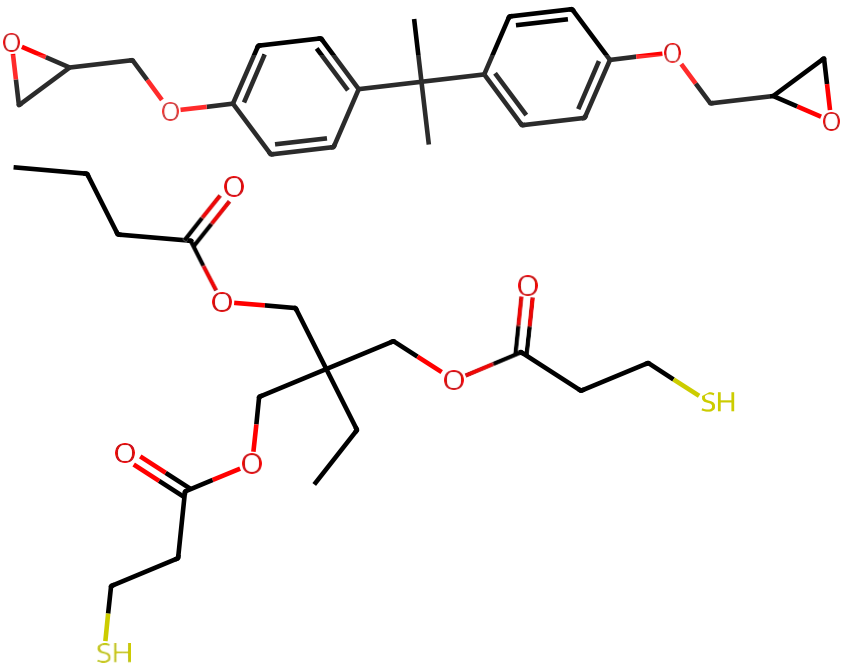

Figure S6: Molecular structures of Example 3

Example S4–S6 showcase representative multi-turn conversations, each followed by the polymer structures predicted by the DeepSeek model.

Example 4

User: What would be a good TSMP combination for my project?

Assistant: What's more important for your TSMP - certain functional groups or specific properties like Tg and Er?

User: I'd rather focus on Tg and Er requirements.

Assistant: I'll need two properties to design your TSMP - what glass transition temperature (Tg) and rubbery modulus (Er) would you like?

User: Generate a thermoset shape memory polymer with Tg = 100°C and Er = 40 MPa.

Assistant: Output Here is a TSMP designed for Tg = 100C and Er = 40Mpa: Monomer 1: C=C(C)C(=O)OCCOCCOCCOCCOCCOCCOCCOCCOCCOC(=O)C(=C)C

Monomer 2: C=C(C)C(=O)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOC(=O)C(=C)C

Molecular structures of these monomers are given below:

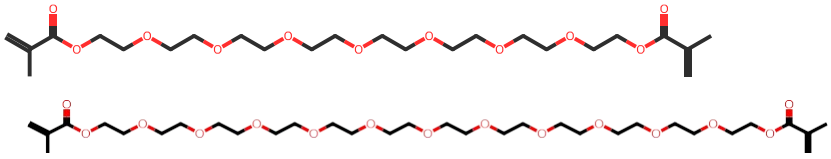

Figure S7: Molecular structures of Example 4

Example S5

User: I'm working on a thermoset shape memory polymer - what should I use? **Assistant:** What's your priority - specific chemical groups or target properties? **User:** I want to design using particular groups while hitting certain property targets.

Assistant: I can create TSMPs using functional groups such as epoxy (C1OC1), imine (NC), vinyl (C=C), or thiol (CCS), while achieving your target Tg and Er specifications.

User: Generate a TSMP with Tg = 100°C and Er = 40 MPa, using functional groups epoxy(C1OC1) and imine (NC).

Assistant:

Output Here's a pair of monomers fulfilling the criteria of Tg = 100C, Er = 40MPa, and containing C1OC1/NC groups:

Monomer1:CC(C/C=CCC(=O)OCC(O)COc3ccc(C(C)(C)c2ccc(OCC1CO1)cc2)cc3)C(=O)OCC(O)COc6ccc(C(C)(C)c5ccc(OCC4CO4)cc5)cc6

Monomer 2: NCCOCCOCCN

Molecular structures of these monomers are given below:

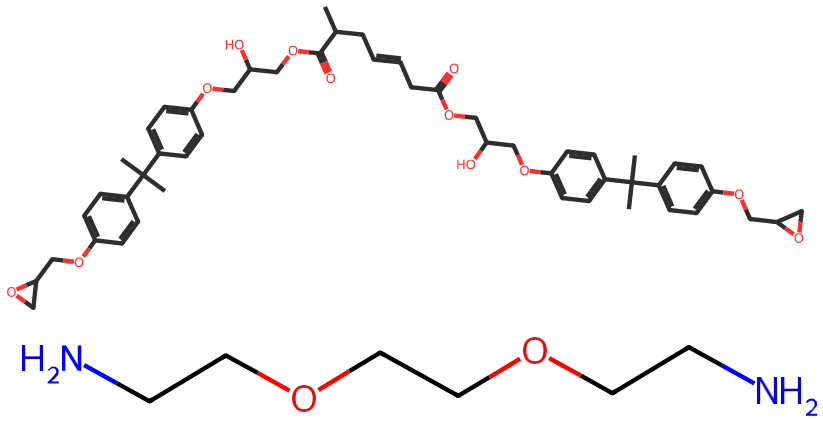

Figure S8: Molecular structures of Example 5

Example S6

User: I'm working on a thermoset shape memory polymer - what should I use? **Assistant:** What's more important for your TSMP - certain functional groups or specific properties like Tg and Er?

User: I'd prefer to work with the chemical groups.

Assistant: I'll need two chemical groups that are more reactive - I have several options. What types of groups are you interested in?

User: Design a thermoset shape memory polymer using monomers that include epoxy(C1OC1) in monomer1 and imine (NC) in monomer2.

Assistant:

Output Here's a pair of monomers containing C1OC1 and NC groups, ideal for TSMP design:

Monomer 1: CC(C)(c2ccc(OCCOCCOCCOCC1CO1)cc2)c4ccc(OCCOCCOCCOCC3CO3)cc4

Monomer2:CC(N)COCC(C)OCC(C)OCC(C)OCC(C)OCC(C)OCC(C)OCCOCCOCCOCCO

CCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCC

OCCOCCOCCOCCOCCOCCOCC(C)OCC(C)OCC(C)OCC(C)OCC(C)N

Molecular structures of these monomers are given below:

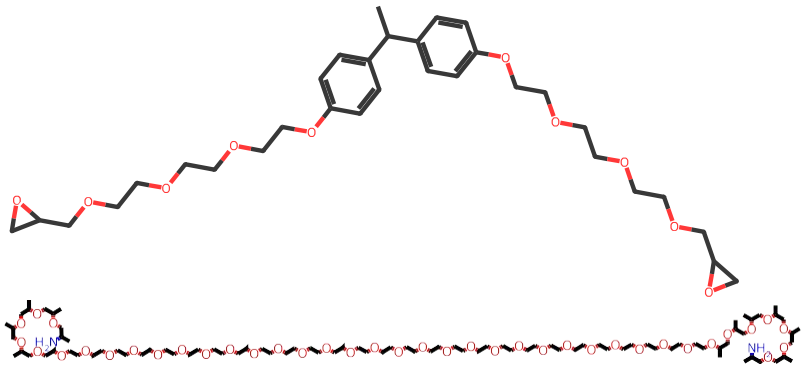

Figure S9: Molecular structures of Example 6

Example S7–S9 illustrate selected examples of multi-turn interactions, accompa- nied by the corresponding polymer candidates generated by the Llama3.2 model.

Example S7

User: I'm working on a thermoset shape memory polymer - what should I use? **Assistant:** To create your TSMP, I need two key properties: the molecular weight and structure. What values would you like for the chemical groups?

User: I want to focus on the chemistry side - functional groups.

Assistant: I can help design TSMPs using available groups such as vinyl (C=C), imine (C=N), or epoxy (C1OC1).

User: Design a thermoset shape memory polymer using monomers that include vinyl(C=C) in monomer1 and hydroxyl (=O) in monomer2.

Assistant:

Output Here's a TSMP that includes the requested functional groups:

Monomer1(containsC=C):C=C(C)C(=O)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOC

COCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOC(=O)C(=C)C

Monomer 2 (contains =O): C=C(C)C(=O)OCCOCCOCCOC(=O)C(=C)C

Molecular structures of these monomers are given below:

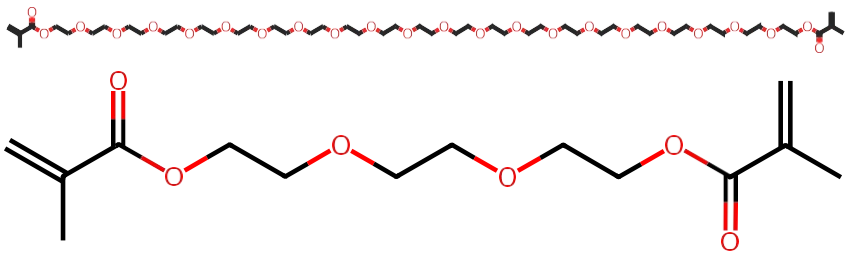

Figure S10: Molecular structures of Example 7

Example S8

User: Hey, I'm looking to make a thermoset shape memory polymer - any ideas? **Assistant:** For your TSMP, are you more interested in the chemistry or the properties? **User:** Properties are more important for my application.

Assistant: We can work with any Tg and Er values - just key in your desired values. **User:** Generate a thermoset shape memory polymer with Tg = 150°C and Er = 30 MPa. **Assistant:**

Output Here's a TSMP that matches your request:

Monomer 1: CC(C)(c2ccc(OCC1CO1)cc2)c4ccc(OCC3CO3)cc4

Monomer2:CC(N)COCC(C)OCC(C)OCC(C)OCCOCCOCCOCCOCCOCC

OCCOCCOCCOCCOCCOCC(C)OCC(C)OCC(C)N

Molecular structures of these monomers are given below:

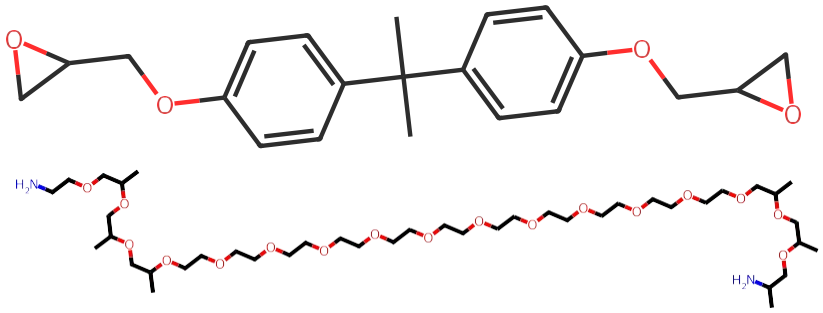

Figure S11: Molecular structures of Example 8

Example S9

User: I'd like to create a shape memory polymer - can you recommend some monomers? **Assistant:** For your TSMP, are you more interested in the chemistry (functional groups) or the properties (Tg/Er)?

User: Let's create a TSMP with specific groups and property requirements.

Assistant: For your TSMP, I can use different functional groups (epoxy (C1OC1), imine (NC), vinyl (C=C), thiol (CCS)) and specify two key properties.

User: Generate a TSMP with Tg = 50°C and Er = 20 MPa, using functional groups vinyl(C=C) and hydroxyl(=O).

Assistant:

Output Here's a valid monomer pair designed with the required functionalities:

Monomer1(rich in C=C): C=C(C)C(=O)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO CCOCCOCCOCCOCCOCCOCCOCCOC(=O)C(=C)C

Monomer 2 (rich in (=O)): C=C(C)C(=O)OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO CCOCCOCCOCCOCCOCCOCCOCCOCCOCCOC(=O)C(=C)C Molecular structures of these monomers are given below:

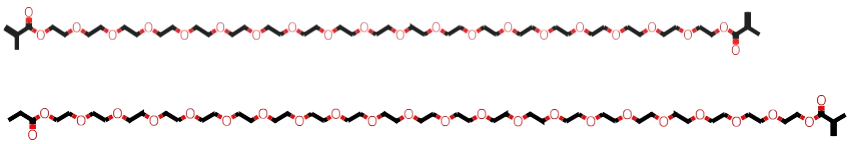

Figure S12: Molecular structures of Example 9

3 Retrieval-based vs Fine-tuning Strategies

Due to space limitations, we do not include the full responses from the both models in the table and instead represent truncated sections with ellipses ('. . . ').

Table S1: Combined qualitative examples and conceptual comparison between the retrieval-based GPT-4o-mini setup and the fine-tuned model.

Query / Feature	**Fine-tuned Prompt2Poly (GPT-4o-mini based)**	**Retrieval-based GPT-4o-mini + attached dataset-based**
Task understanding	Understands polymer-design objectives and generates new TSMPs aligned with target properties.	References or summarizes file content; lacks deep interpretation of design intent.
Adaptivity	Asks follow-up questions when in- put is incomplete; supports multi- turn reasoning.	Cannot clarify missing property or group information; single-turn interaction.
Output type	Generates valid, property-aligned SMILES sequences.	Produces descriptive text or data references or SMILES from the attached file.
Interaction style generation.	Context-aware, iterative dialogue	Static, one-step responses.
Please generate a TSMP with T_g=40C and E_r=50MPa.	Here's a monomer pair that includes the required func- tions and meets the tar- get properties: Monomer 1: CC(C)(c2ccc(OCCOCCOC COC-COCC2)CC1CO1)c4ccc..... Monomer 2: NCC-NCCN(CCNCCC(NCCN)....	To design a thermoset shape memory polymer (TSMP) with a glass transition temperature (T_g) of 40°C and a storage modu- lus (E_r) of 50 MPa, we need to consider the selection of appro- priate monomers and crosslink- ing agents. Here are the steps and considerations for generating such a polymer.....
Generate a TSMP including T_g=100C and E_r=100Mpa with Epoxy group and Imine	Here's a monomer pair that includes the required functions and meets the target properties: Monomer 1: O=C(OCC1CO1)C3CC2O C2CC3C(=O)OCC4CO4 Monomer 2: NCCOCCOCCN	To create a thermoset shape mem- ory polymer (TSMP) system tar- geting a glass transition temper- ature (T_g) of 140 °C and a re- covery stress (E_r) of 40 MPa with C=C and CCS functional- ities, I propose the following novel monomer pair: **Monomer 1:** - Structure: C=C attached to rigid aromatic components to enhance T_g. - SMILES: C=CC1=CC2=C(C=.....

4 Computational Resources and Environmental Impact

Table S2: Estimated Computational Cost and Energy Use for Fine-Tuning.

Model	Platform	GPU Type / API	# GPUs	GPU-Hours	Power Draw (kW)	PUE	Energy (kWh)	CO_2 (kg eq)	Cost (USD)
GPT-4o-mini	OpenAI API (cloud)	Proprietary	—	—	—	—	—	—	$13.66
DeepSeek R1	LONI Cluster (Linux)	4 × A100 80 GB PCIe	4	2.13	0.30	1.2	≈ 0.77	≈ 0.26	Free allocation
LLaMA-3.2	LONI Cluster (Linux)	4 × A100 80 GB PCIe	4	0.47	0.30	1.2	≈ 0.17	≈ 0.05	Free allocation

5 Comparative Overview of Generative Frameworks

Table S3: Conceptual Comparison Between VAE-, GAN-, and LLM-Based Models.

Model Type	Input / Conditioning	Property Control	Novelty	Diversity	Interpretability	Key Limitation
VAE	Encoded molecular features or latent vectors	Indirect; requires conditional head	Moderate to high—depends on latent regularization	Moderate; may collapse with strong regularization	Moderate (latent arithmetic partially interpretable)	Limited control over multiple properties; restricted flexibility
GAN	Random noise vector	Indirect; guided through discriminator's feedback	Low to Moderate when mode collapse is avoided	Variable; prone to mode collapse and instability	Low	Training instability; lacks text or reasoning-based condition- ing
LLM (Prompt2Poly)	Natural-language prompts with property and group tokens	Direct; interprets textual constraints	High—enabled by language-driven combinatorial reasoning	Moderate to high; supports constraint-guided structural diversification	High (prompt-based explainability and token-level reasoning)	Requires chemistry-aware prompt design; higher computational demand