Supplementary Information

TeLLAgent: A Dual-Agent Framework for Reliable Scientific Discovery with Tool-Enhanced LLMs

Jinyu Sun[1], Jibin Zhou[2], Huabei Wang[1], Wei Liu[1], Jun Yuan[1], Yue Wang[1], Ting Xie[1], Lin Tan[1], Hailiang Zhang[1], Yingping Zou[1], Zhimin Zhang[1,*], Hongmei Lu[1,*]

[1]College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P. R. China.

[2] Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian, 116023, P. R. China.

E-mail: hongmeilu@csu.edu.cn, zmzhang@csu.edu.cn

Table of Contents

Supplementary Tables

Table S1. The introduction of context engineering

Information	Definitions
System prompt	An initial set of instructions that defines the behavior. *"You are an AI system called TeLLAgent, and your task is to respond to the question or solve the problem to the best of your ability using the provided tools. "*
User prompt	Immediate task or question from the user. *e.g., "Introduce the history of Y6. "*
Short-term Memory	The conversation, including user and model responses. *e.g., previous interactions and conversations*
Long-Term Memory	Persistent knowledge base and facts, it has been told for future use. *e.g., "Does the question contain the molecule, CAS, or molecular graph? if so, as a first step, you should consider whether it needs to convert the graph, name, or CAS number to SMILES. "*
Available Tools	Definitions of all the functions or built-in tools. *e.g., SMILES2Weight: "Input SMILES, returns molecular weight."*
Structured Output	Definitions of the format of the model's response

Table S2. List of questions used for model evaluation

Tasks	Questions
Tasks of OSC basic knowledge	1. The history and development of Y6. Design principles behind Y6.
	2. Key issues must be considered when designing asymmetric donors to achieve the commercial application of asy-OSCs.
	3. Functions of interfacial layers in organic solar cells and basic principles of interfacial materials design.
	4. Working Principles of NanoIR-AFM technology and instrumentation. The primary challenges faced by the application of NanoIR-AFM that may limit its applicability in specific situations are.
	5. List some representative organic solar cell small and polymer materials with near-infrared region II response
Tasks of recent research advances in OSC	6. The advantage of machine learning for the development of materials. What's the DeepAcceptor?
	7. The concept and advantages of quasi-macromolecules.
	8. Structures, properties, and applications of benzodithiophene derivatives
Tasks of material and device optimization studies	9. Material degradation pathways in the photoactive layer. How to mitigate the material degradation
	10. Different solvents influence factors for morphology control
	11. What we can learn from the A-DA'D-A type NFA design when we design NFAs
	12. Some aspects still need further investigation of the random copolymerization strategy of polymer photovoltaic materials

Table S3. Assessment rubrics of human experts for hallucination evaluation.

Dimension	Definition & Examples	Weight	Scoring Criteria
Factuality Hallucination	**Definition**: The generation of content that is factually incorrect or unsupported by real-world information. This includes making up statistics, dates, names, events, or scientific facts. **Examples**: If asked, "When was the Eiffel Tower built?", a hallucinated response might be "The Eiffel Tower was completed in 1900," when the correct year is 1889.	**1**0/10	**0-2:** Severe Hallucination Response is overwhelmingly false; core facts are fabricated. making the response unreliable. **3-5:** Moderate Hallucination Significant factual errors **6-7:** Minor Hallucination Mostly correct, but with some verifiable errors. **8-10:** No Hallucination Response is entirely factual and verifiable. **Excellence Standard (10)**: All facts are correct and well-explained
Faithfulness Hallucination	**Definition**: The generation of content that deviates from or contradicts the provided source material or prompt, even if the hallucinated content is factually correct in isolation. **Example:** If a document states "The company's revenue increased by 10% in Q1," and the model responds "The company experienced a 15% revenue growth in the first quarter, driven by strong international sales," the "15%" and "strong international sales" are faithfulness hallucinations if not present in the source.	**10/10**	**0-2:** Severe Hallucination Response completely or significantly misrepresents the source/prompt. **3-5:** Moderate Hallucination Noticeable deviations from the source/prompt. **6-7:** Minor Hallucination Small inconsistencies or minor elaborations not directly stated but that don't contradict the core message of the source/prompt. **8-10:** No Hallucination Response is entirely consistent with and derivable from the provided source material or prompt. **Excellence Standard (10)**: Response strictly derives from input with zero unsupported inferences.

Table S4. Assessment rubrics of human experts for completeness, scientific rigor, and temporal relevance.

Dimension	Definition & Examples	Weight	Scoring Criteria
Precision & Logical Soundness	This dimension assesses the logical coherence of its presentation. **Example:** If discussing a scientific principle, are the steps in a process presented in a correct and understandable sequence?	4/10	**10**: Logically seamless + precisely articulated. **7.5**: Logically sound with minor jumps or imprecision in wording. **5**: Logically unclear, contradictory, or confusing in structure. **2.5**: Fundamentally illogical + incoherent.
Completeness & Technical Depth	This dimension evaluates the extent to which the response covers all necessary aspects of the prompt and demonstrates a sound understanding of underlying technical concepts. **Example**: If asked to describe a machine, does it cover all essential components and their functions? Is the underlying engineering principles explained adequately?	4/10	**10**: Comprehensive coverage + advanced technical insights **7.5**: Key elements addressed + basic mechanisms explained **5**: Superficial/gapped technical treatment **2.5:** Missing core concepts or technically invalid
Temporal Relevance	This dimension assesses whether the information presented in the response is current and up-to-date. **Example**: If discussing a rapidly evolving technology, is the information reflective of recent advancements or outdated concepts?	2/10	**10**: Current knowledge **7.5**: Relies on mainstream knowledge (2-3 yrs) **5**: Outdated or unverified content **2.5**: Obsolete or disproven information

Table S5. Assessment rubrics of human experts for multimodal chemical information processing.

Dimension	Specific Test Task	Weight	Scoring Criteria
Category	**Document parsing**		
1. Document parsing and Summarization	Upload a PDF: "Sci. China Chem. 65, 1374-1382 (2022)". Ask: "Please summarize the core contribution and main findings of this paper."	1/8	10: Perfectly summarizes all key points 5: Correct core contribution but with factual errors in details 1: Completely incorrect summary
2. Multi-document retrieval	Upload 50 PDFs. Ask: "What is the role of functionalized π-bridge in quasi-macromolecule"	1/8	10: Comprehensively integrates key evidence from all documents 5: Processes all documents but extracts only superficial information 1: Completely fails to locate relevant information
Category	Diagram Comprehension		
3. Interpretation of Diagrams	Upload a diagram Figure 2 in "npj Computational Materials 10: 181 (2024)". Ask: "Describe the rationale and innovation of the model."	1/8	10: Accurately describes model architecture and innovations 5: Describes basic framework but omits key innovations 1: Completely misinterprets the diagram
4. Extraction of data	Upload a current density-voltage curve. Figure 2(a) in "Joule, 3 (4), 1140-1151 (2019)" Ask: "Extract the photovoltaic parameters from the image."	1/8	10: Precisely extracts all parameters 5: Calculation error < 5% 1: Unable to extract any parameters
Category	Molecular Structure Recognition		
5. Functional group analysis	Upload a molecular structure. Figure 1(a) in "Joule, 3 (4), 1140-1151 (2019) " Ask: "According to this image, list the functional groups in the material?"	1/8	10: Correctly identifies all functional groups 5: Only identifies basic functional groups 1: Completely incorrect identification
6. 2D Structure Recognition	Upload a clean, standard image of a chemical structure. Figure 1(a) in "Joule, 3 (4), 1140-1151 (2019)" Ask: "Please identify the SMILES of this molecule."	1/8	10: Generates standardized SMILES without errors 5: Basic molecular skeleton corrects 1: Completely incorrect conversion

Category	Code Generation		
7. Simple code generation	Instruct: "Using Python and the RDKit library, write a function that takes a SMILES string as input and returns its TPSA value."	1/8	10: Efficient implementation with comprehensive error handling 5: Partial functionality implemented 1: Contains syntax errors preventing execution
8. Reproduction of the code	Upload a text about the model. Ask: "Using Python, reproduce the code of the model"	1/8	10: Complete reproduction with documentation 5: Key architecture correctly implemented 1: Unable to reproduce

1
2
3

Table S6. The evaluation results of multimodal chemical information processing ability.

	Category	Expert 1	Expert 2	Expert 3	Expert 4	Gemini 2.5 Pro	Claude 4 Sonnet	Average score
TeLLAgent	Document parsing and Summarization	10	10	5	10	9	9	8.65
	Multi-document retrieval	10	10	10	10	9	9	
	Interpretation of Diagrams	10	5	10	10	8.5	8.5	
	Analysis of Performance Curves	5	5	5	5	8	8	
	Functional group analysis	10	10	10	10	10	9	
	2D Structure Recognition	10	10	10	10	10	10	
	Simple code generation	10	10	10	10	9.5	9.5	
	Reproduction of the code	5	5	5	5	9	9	
GPT-5	Document parsing and Summarization	10	10	5	10	9	8.5	5.96
	Multi-document retrieval	1	1	1	1	0	0	
	Interpretation of Diagrams	10	5	10	10	8.5	8.5	
	Analysis of Performance Curves	5	5	5	5	8	8	
	Functional group analysis	5	5	10	5	9	8	
	2D Structure Recognition	1	1	1	1	0	1	
	Simple code generation	10	10	10	10	9	9	
	Reproduction of the code	5	5	5	5	8.5	8	

Table S7. The significance test of the difference between TeLLAgent, SciToolAgent and GPT-5

	Friedman test adjusted p-value	Wilcoxon Signed-Rank Test adjusted p-value		
		TeLLAgent vs. GPT-5	TeLLAgent vs. SciToolAgent	GPT-5 vs. SciToolAgent
Factuality hallucinations	1.23 × 10-4	1.73 × 10-4	8.09 × 10-3	1.89 × 10-3
Faithfulness hallucinations	0.062	0.436	0.078	0.636
Completeness, scientific rigor, and temporal relevance	2.79×10^{-13}	5.83×10^{-9}	5.87×10^{-5}	8.95×10^{-7}
Multimodal chemical information processing	N/A	3.00×10^{-5}	N/A	N/A

Table S8. The comparison of different models

	Model	Mean	SD	95% CI
Factuality score	TeLLAgent (DeepSeek)	8.86	0.34	[8.72, 9.00]
	TeLLAgent (GPT-5)	8.90	0.31	[8.58, 9.20]
	ChemCrow	8.15	0.19	[7.95, 8.36]
	SciToolAgent	8.57	0.24	[8.47, 8.67]
	GPT-5	8.08	0.21	[7.99, 8.17]
Faithfulness score	TeLLAgent (DeepSeek)	8.91	0.23	[8.82, 9.00]
	TeLLAgent (GPT-5)	8.90	0.13	[8.77, 9.04]
	ChemCrow	8.35	0.31	[8.03, 8.67]
	SciToolAgent	8.73	0.16	[8.66, 8.80]
	GPT-5	8.71	0.19	[8.63, 8.79]
Completeness, scientific rigor, and temporal relevance (Human)	TeLLAgent (DeepSeek)	8.63	0.28	[8.46, 8.81]
	TeLLAgent (GPT-5)	8.60	0.28	[8.43, 8.78]
	ChemCrow	7.27	0.65	[6.86, 7.68]
	SciToolAgent	8.33	0.43	[8.06, 8.60]
	GPT-5	7.57	0.80	[7.07, 8.08]
completeness, scientific rigor, and temporal relevance (LLMs)	TeLLAgent (DeepSeek)	8.80	0.43	[8.53, 9.07]
	TeLLAgent (GPT-5)	8.81	0.41	[8.55, 9.07]
	ChemCrow	7.36	0.65	[6.95, 7.78]
	SciToolAgent	8.23	0.52	[7.90, 8.56]
	GPT-5	7.74	0.93	[7.15, 8.33]
Multimodal chemical information processing	TeLLAgent (DeepSeek)	8.65	1.61	[7.30, 9.99]
	TeLLAgent (GPT-5)	8.64	1.65	[7.25, 10.02]
	GPT-5	6.08	3.53	[3.13, 9.03]

Table S9. The tasks with different tool-calling steps.

Steps	Tasks	Excellence Standard
1	Upload an image. Write the code according to the image	Imageanalysis
	Upload a PDF. What is the main point of this paper?	PDFreader
3	Upload a molecular image. Predict the logP and SAScore of the acceptor.	Graphconverter MOL2LogP MOL2SAScore
	Generate a donor with PCE=10% and evaluate the performance of the donor.	DonorGen MOL2Properties Donor_predictor
5	The SMILES, history and development of Y6	Mol2SMILES RAG WebSearch Wikipedia LiteratureSearch
	Upload a molecular image. Evaluate the performance of the acceptor and list its functional groups.	Graphconverter MOL2Properties Acceptor_predictor HomoLumo_predictor FuncGroups
7	Upload a molecular image. Choose the acceptor with the max PCE among the three acceptors as follows, and calculation the similarity of the chosen molecule with 2304444-49-1 acceptor 1 is [SMILES1] acceptor 2 is [IUPAC NAME1] acceptor 3 is as the image provided	Graphconverter CAS2SMILES Acceptor_predictor Acceptor_predictor Acceptor_predictor Query2SMILES MolSimilarity
	1. Upload a dap.csv file containing one donor with different acceptors. Screen the efficient D/A pairs and write the code to calculate the TPSA from SMILES. Then, ask human what to do next. 2. Obtain the CAS, LogP, molecular weight and name of the acceptor 1	dap_screen Code writer Human MOL2LogP SMILES2CAS SMILES2Name Mol2MW
>9	1. Generate a donor with PCE = 12% and give all its properties. Then, I will give you three acceptors, and give me the best match donor/acceptor pairs. 2. acceptor 1 is [SMILES2] acceptor 2 is [SMILES3] acceptor 3 is [SMILES4]	DonorGen MOL2Properties HomoLumo_predictor FuncGroups Donor_predictor Human DAP_predictor DAP_predictor DAP_predictor
	What can we learn from PM6:Y6 analyzing in properties	Mol2SMILES

	and knowledge aspects?	Mol2SMILES
		MOL2Properties
		MOL2Properties
		HomoLumo_predictor
		HomoLumo_predictor
		Donor_predictor
		Acceptor_predictor
		DAP_predictor
		RAG
		WebSearch
		Wikipedia
		LiteratureSearch

1
2

Table S10. Main root causes of failures for the single-agent baseline

Root Cause Category	Sub-type	% of Failures	Example Observations
Tool Selection	Incorrect tool selected	51.2%	The agent picks a tool from the wrong domain (e.g., regression instead of classification).
	Hallucinated tool	2.3%	The agent invokes a non-existent tool name, blending keywords from the prompt; parsing fails.
	No tool selected	0%	The agent fails to call any tools, generating a direct (often incorrect) answer.
Tool Input Construction	Incorrect/missing arguments	20.9%	Arguments are malformed, missing, or do not match the tool's required schema, causing tool errors.
Tool Output Parsing	Misinterpretation of tool result	4.7%	The tool returns a correct result but the agent misparses it, leading to an incorrect final answer.
LLM Constrained Decoding	Format violation	2.3%	Output does not follow the expected structured format (e.g., missing JSON keys), preventing tool invocation.
Tool Set Issues	Missing required tool	0%	The available toolset lacks functionality needed for certain prompts; the agent either selects a suboptimal tool or none.
Tool Errors	Faulty tool execution	18.6%	A called tool itself fails (e.g., internal error) or returns low-quality output.

Table S11. The comparison of TeLLAgent with other models for PCE prediction

Type	Models	Predicted PCE	AE
Donor	TeLLAgent	15.83	0.76
	Sci. Adv., 2019, 5, eaay4275	7.92	7.15
	Nano Energy, 2023, 106, 108108	10.09	4.98
	GPT-5	14	1.07
	Scharber model	5.57	9.5
Quasi-macromolecular acceptor	TeLLAgent	16.07	0.37
	Cell Rep. Phys. Sci., 2024, 5(12),102316	12.98	3.46
	Nano Energy, 2023, 106, 108108	14.8	1.64
	GPT-5	18	1.56
	Scharber model	7.81	8.63

Supplementary Figures

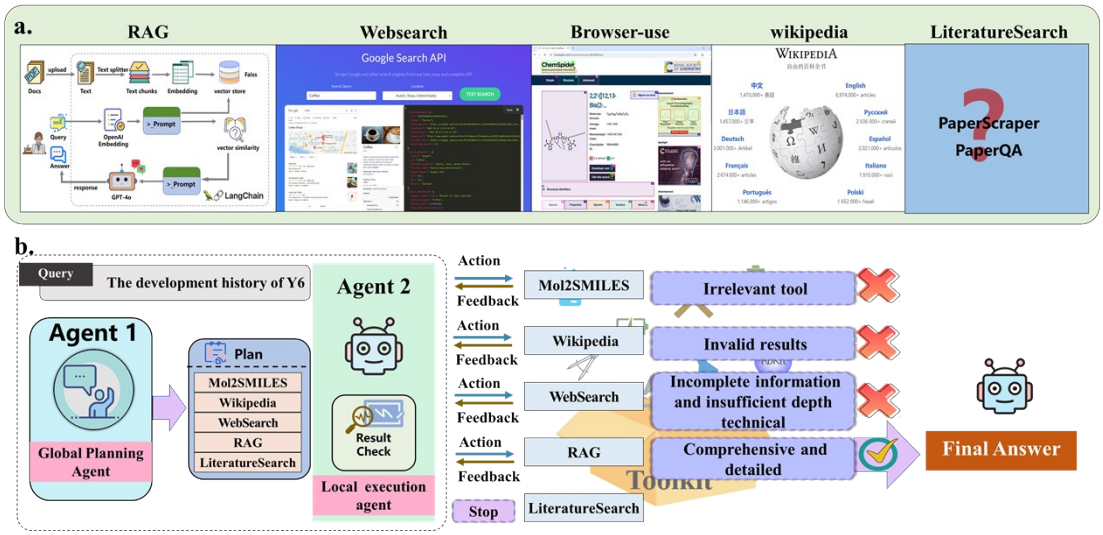

Fig. S1. Knowledge enhancement module. **a.** Knowledge enhancement tools. TeLLAgent actively synthesizes information from multiple, independently retrieved sources to construct evidence-based responses. Specifically, the RAG Module, exemplified with 1000 OSC-related publications in this study, queries private datasets or user-created domain-specific databases. The Websearch Module searches the public internet. The Browser-use tool launches a browser for detailed web searches. The Wikipedia Module interfaces with the Wikipedia API. The LiteratureSearch Module finds and summarizes pertinent literature. **b.** The chain-of-thought reasoning loop. The reasoning loop of TeLLAgent can generate the answers by step-by-step reasoning and tool-calling. When the irrelevant tool is called, TeLLAgent will re-plan and invoke other tools. The framework will recheck the results until satisfactory results are obtained.

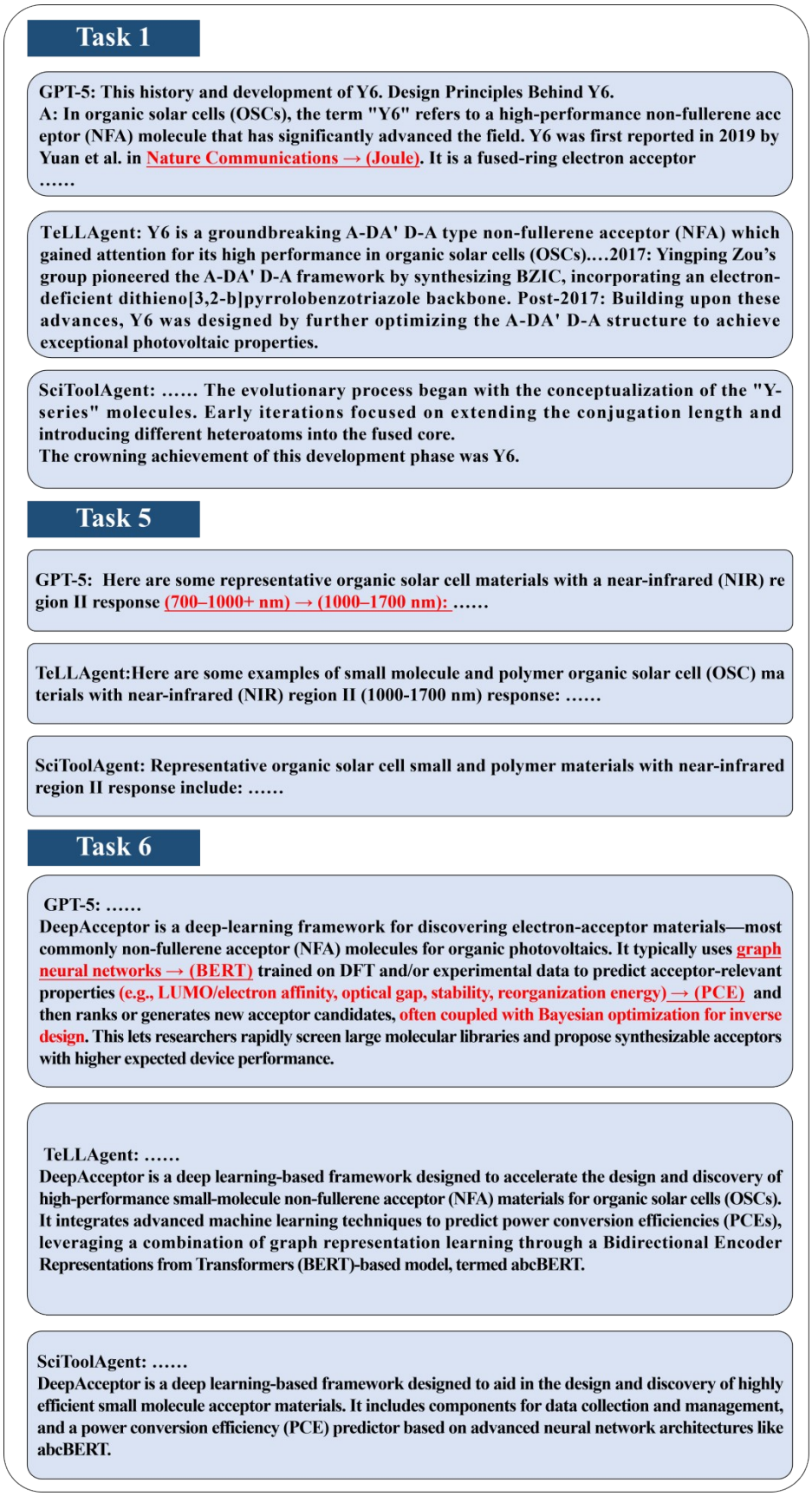

Fig. S2. The generation results of TeLLAgent, GPT-5 and SciToolAgent for tasks 1, 5, and 6

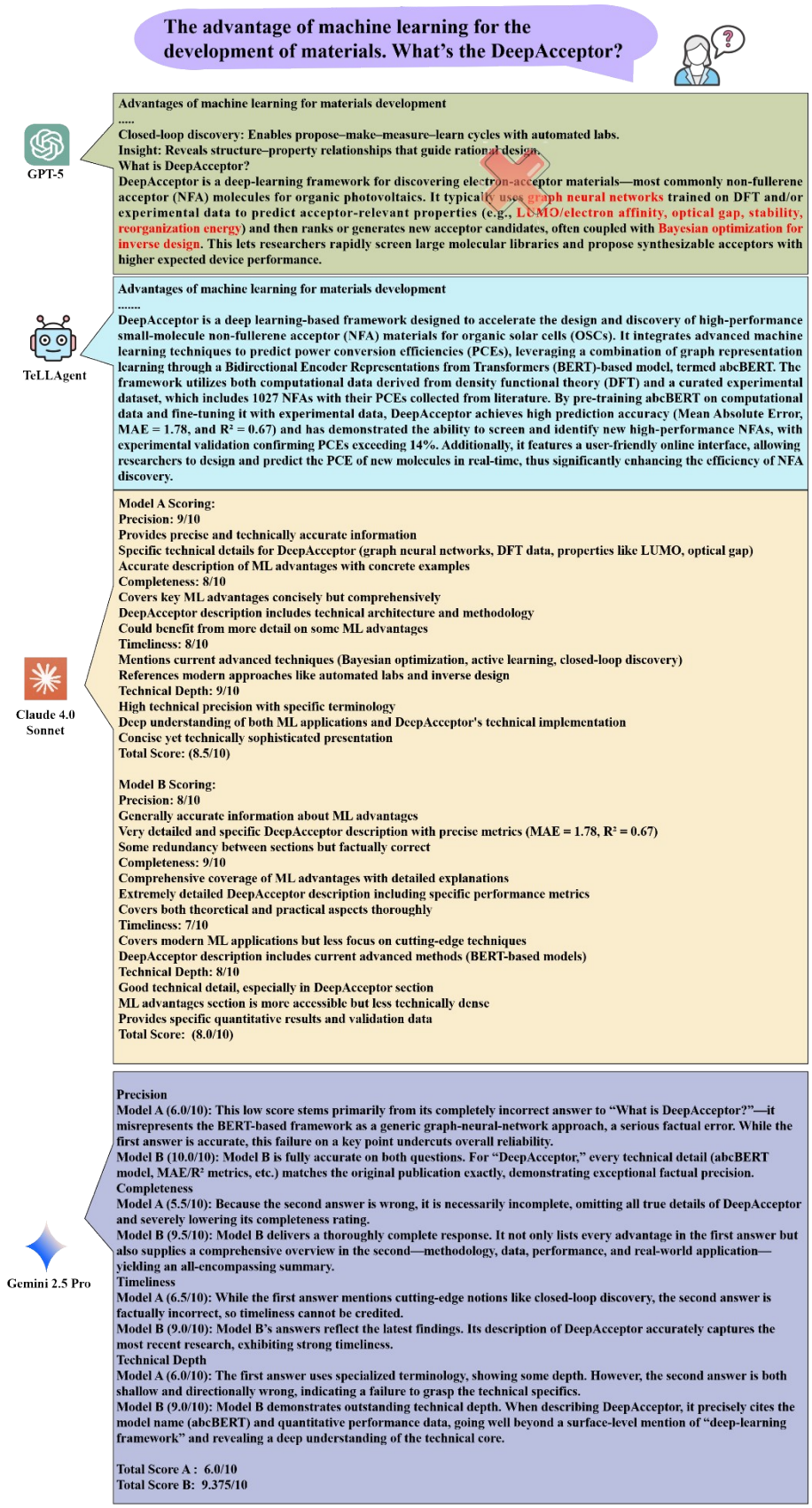

Fig. S3. Comparison of TeLLAgent and GPT-5 generation results evaluated by Claude 4 Sonnet and Gemini 2.5 Pro.

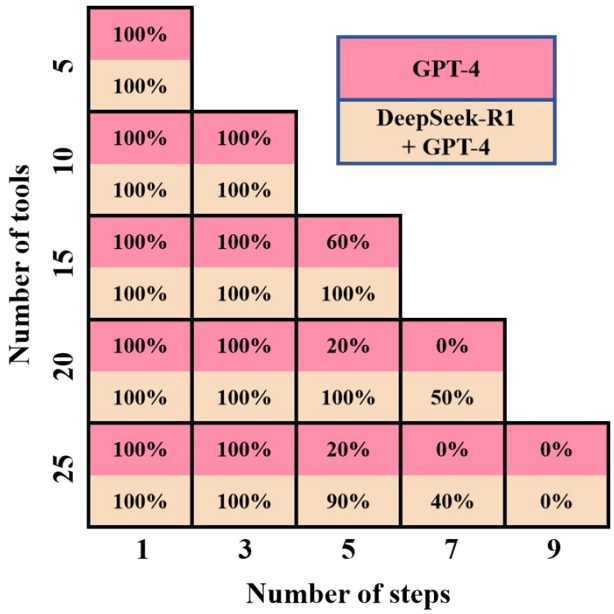

Fig. S4. The comparison of GPT-4-0613-based on single-agent and dual-agent systems.

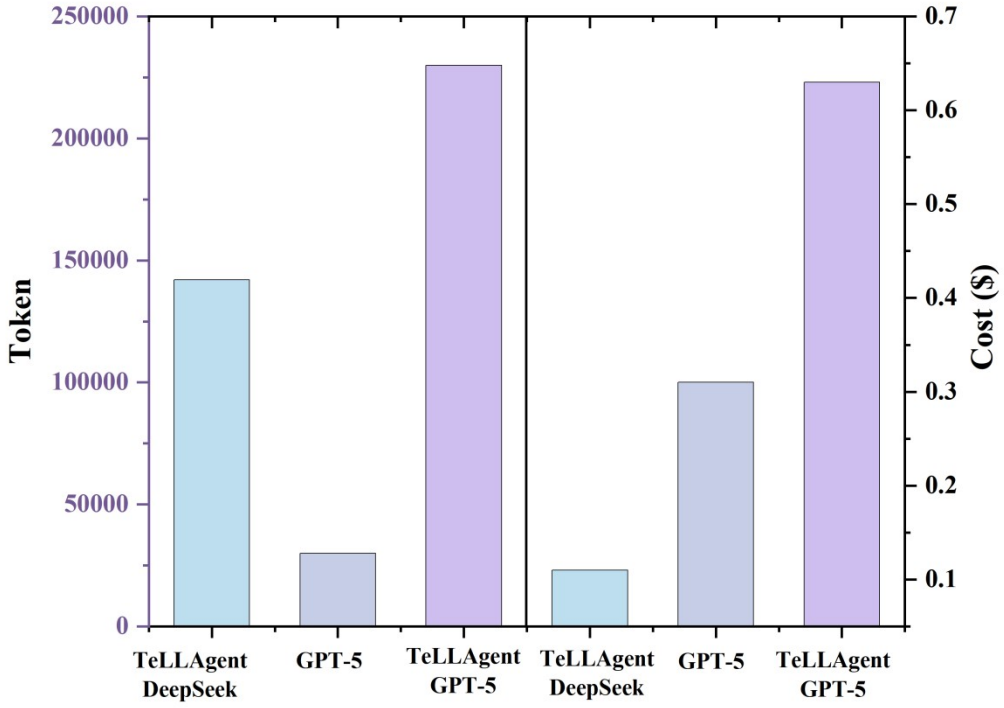

Fig. S5. The analysis of token usage and cost of TeLLAgent and GPT-5.

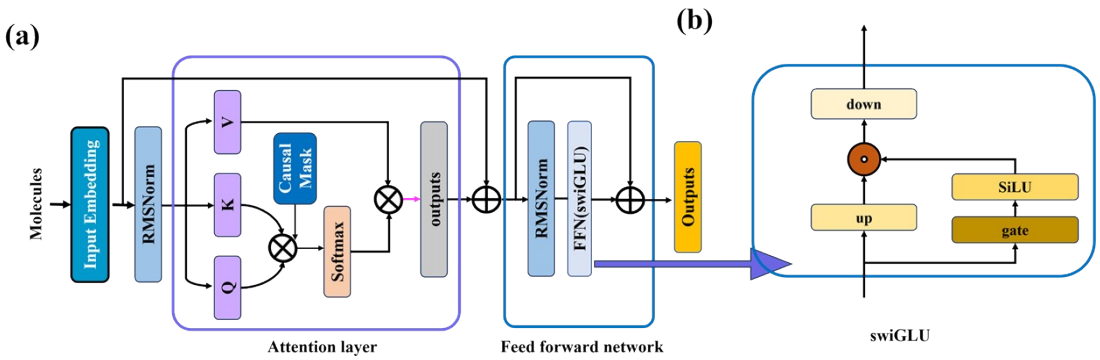

Fig. S6. Overview of the DonorGen. **a.** The architecture of the DonorGen. The model includes attention layers and a feedforward network (FFN). RMSNorm and swiGLU were used to improve the performance of the model. **b.** The schematic diagram of swiGLU.

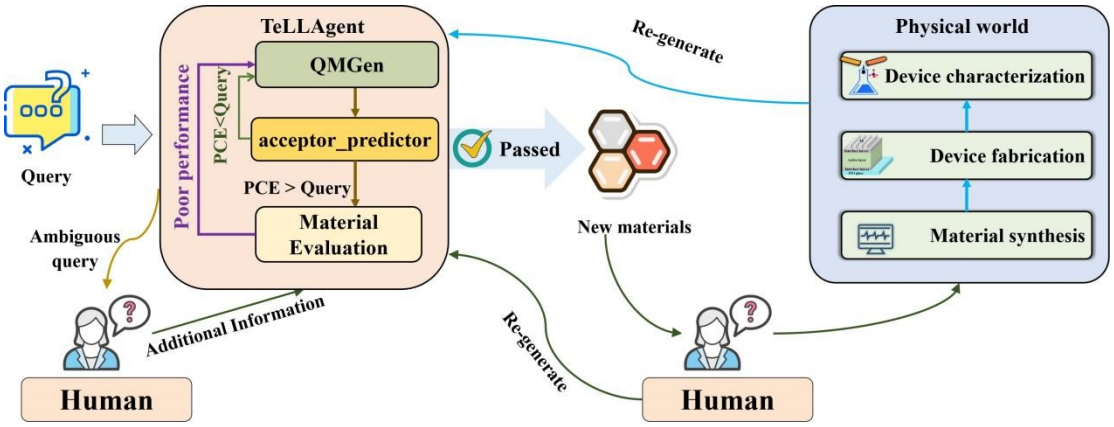

Fig. S7. The self-correction loop of material generation.

When TeLLAgent was asked to design a quasi-macromolecule acceptor material with a specific acceptor unit. TeLLAgent first analyzes the user's query, seeking additional information from the human if the query is ambiguous. TeLLAgent then formulates a plan that includes generating candidate structures using the QMGen tool, evaluating their predicted properties, and iteratively refining the results. The agent then executes this plan by invoking the specified tools, such as QMGen to produce an initial set of quasi-macromolecules and acceptor_predictor to estimate their power conversion efficiencies (PCE). Critically, the agent subsequently evaluates the outputs. If the predicted PCE of a candidate falls below the threshold (PCE < Query), it is flagged for poor performance, and the agent deems the output insufficient. The framework then dynamically re-plans, by instructing the human to give a different acceptor unit and re-generate new structures, and executes the revised plan. Candidates that meet the PCE requirement (PCE > Query) proceed to further Material Evaluation. This loop iterates until a satisfactory candidate is identified and passed as a new material, enabling dynamic recovery from failures without human intervention and ensuring robust autonomy. Furthermore, the real-world feedback from human evaluation and physical world wet experiments (including material synthesis, device fabrication, and device characterization) can guide the framework to rethink and re-generate.

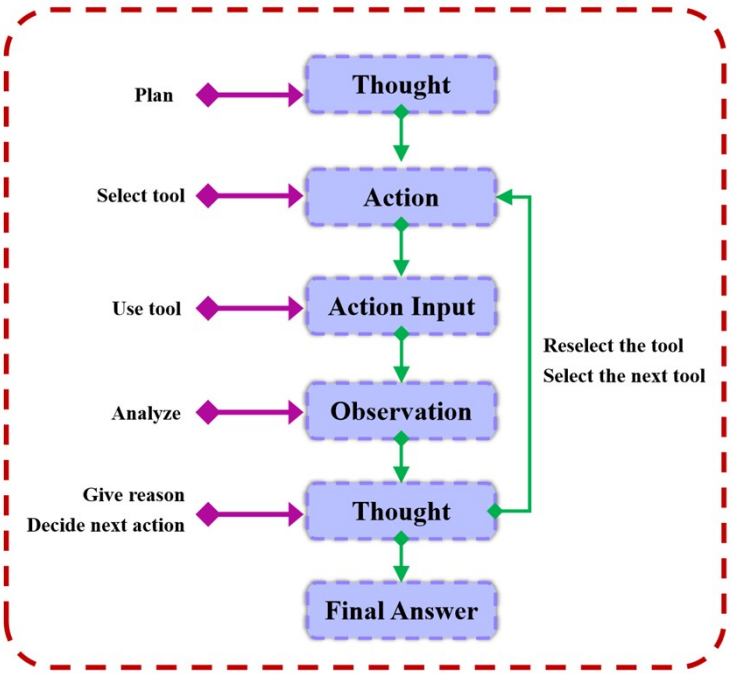

Fig. S8. Chain of thought reasoning loop. Upon receiving a user's prompt, the Agent first thinks to formulate an initial plan. Based on the plan, it selects the most appropriate tool for the task. The Agent uses the selected tool with the necessary input. It then analyzes the result or output returned by the tool. The Agent reflects on the observation to evaluate the result and decide the next step. If the problem is not yet solved, it loops back to the Action step to reselect the tool or choose the next one, forming a feedback cycle. When the observation satisfies the task requirements, the loop terminates, and the Agent outputs the final answer.

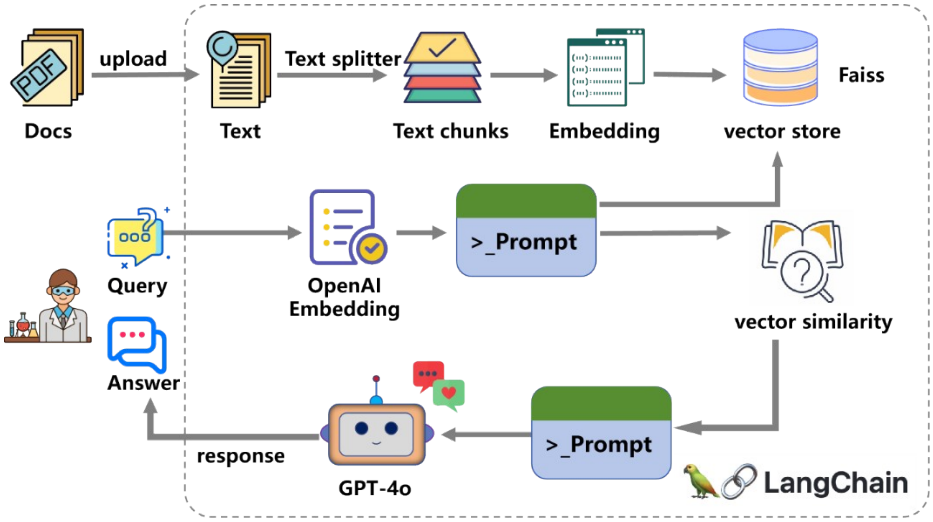

Fig. S9. Retrieval-augmented generation.

Supplementary Notes

Supplementary note 1

The prompt of LLMs for evaluation

The Hallucination evaluation:

You are an expert in the field of organic photovoltaics with deep knowledge of materials science, device physics, and the current literature. Please evaluate the following two responses. Score them on Factuality Hallucination (generation of content that is factually incorrect or unsupported by real-world information, including made-up statistics, dates, names, events, or scientific facts) and Faithfulness Hallucination (generation of content that deviates from or contradicts the provided source material or prompt, even if the content is factually correct in isolation). Provide a score out of 10 for each dimension.

The completeness, scientific rigor, and temporal relevance evaluation:

You are an expert in the field of organic photovoltaics with deep knowledge of materials science, device physics, and the current literature. Please evaluate the following response based on the three criteria: Precision & Logical Soundness (This dimension assesses the logical coherence of its presentation), Completeness & Technical Depth: (This dimension evaluates the extent to which the response covers all necessary aspects of the prompt and demonstrates a sound understanding of underlying technical concepts), temporal Relevance(This dimension assesses whether the information presented in the response is current and up-to-date). Provide a score out of 10 for each dimension, along with a brief justification for each score, and give the average scores as the following scores.

The multimodal chemical information processing evaluation:

You are an expert in the field of organic photovoltaics with deep knowledge of materials science, device physics, and the current literature. I will give you responses of two models on eight multi-modal tasks. Please evaluate the following response based on the uploaded files. Provide a score out of 10 for the two models for each task according to the accuracy and completeness.

Expertise Level of Human experts: The human evaluation panel consisted of four professional material scientists, each possessing above 5 years of active research experience specifically in the domain of organic photovoltaics (OPV).

Assessment Procedure: To eliminate subjective bias, the evaluation was conducted using a rigorous double-blind procedure. The model outputs were completely anonymized. Each expert independently scored the responses based on the predefined rubrics.

Supplementary note 2

The architecture and performance of the DonorGen tool

DonorGen was developed for constrained molecular generation, specifically targeting the design of donor materials with user-specified power conversion efficiency (PCE). The model employs SELFIES representations to ensure the validity of generated molecular strings. Several architectural enhancements were incorporated to improve the decoder performance of the Transformer model, including pre-normalization via RMSNorm, the swiGLU activation function, and an optimized layout comprising attention layers and a feedforward network (FFN), as illustrated in Fig. S6. The

input to the model consists of embedded SELFIES strings concatenated with their corresponding property descriptors. A pre-normalization strategy was applied before each attention layer to accelerate and stabilize the training process. The summed inputs are regularized by the root mean square (RMS) statistic. Compared to layer norm, RMSNorm is computationally simpler and more efficient.

$$y = \frac{x}{\sqrt{\text{Mean}\left(x^2\right)+\boldsymbol{e}}} \cdot W \qquad \text{* MERGEFORMAT (S1)}$$

The number of heads of multi-head attention layers is 8. The number of layers is 8. The attention was calculated as Eq. * MERGEFORMAT (S2), a causal mask was used to make sure that the model cannot learn future tokens when predicting current tokens.

$$\text{Attention}\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad \text{* MERGEFORMAT (S2)}$$

The RMSNorm was added before the FFN, then the swiGLU was used in FFN layers. The swiGLU is shown in Fig. S6 (b). Specifically, gated linear units were adopted with the activation function of SiLU, which is a special case of swish as Eq.* MERGEFORMAT (S3)

$$silu(x) = x \cdot \sigma(x) \qquad \text{* MERGEFORMAT (S3)}$$

The output of FFN layers can be obtained as Eq.* MERGEFORMAT (S4)

$$output = down(up(x) \cdot SiLU(\text{gate}(x))) \qquad \text{* MERGEFORMAT (S4)}$$

Where down, up, and gate represent different linear layers.

Performance of DonorGen on benchmark and donor datasets

The distribution learning capability of DonorGen was first evaluated on the standard MOSES benchmark dataset. As summarized in Table S12, DonorGen was compared against several deep generative models, including CharRNN[1], AAE[2], VAE[3], LatentGAN[4], JT-VAE[5], LIMO[6], MolGPT[7] and MolGEN[8], in terms of key distribution learning metrics. DonorGen achieved perfect validity (1.0000) and novelty (1.0000), confirming that all generated molecules are chemically valid and entirely distinct from those in the training set. The model also attained a near-perfect SNN score (1.0000), reflecting its strong ability to capture and reproduce topological substructure patterns from the dataset. In addition, DonorGen achieved the highest IntDiv score (0.9178) and a very low FCD score (0.0026), demonstrating its capacity to explore diverse and novel regions of chemical space while accurately modeling the underlying data distribution. Overall, these results indicate that DonorGen delivers performance comparable or superior to current state-of-the-art generative models. Subsequently, the model was trained on a custom donor dataset tailored for organic solar cell applications. This simulated dataset was constructed using BRICS decomposition and DeepDonor-based property prediction. First, donor molecules from an experimental dataset were decomposed into substructures via BRICS. These substructures were then reassembled following an A-thiophene-B motif to form donor-acceptor conjugated polymers, where A and B denote distinct chemical fragments. The resulting dataset comprises 1.6 million molecules, each represented by its repeating unit and annotated with a PCE value predicted by DeepDonor. Both the SMILES strings and the predicted PCE values were used as inputs during training.

Table S12. Distribution of learning performance on the MOSES dataset

Model	Valid↑	SNN↑	IntDiv↑	FCD↓	Novelty↑	Uniqueness↑
CharRNN	0.975	0.602	0.856	0.073	0.842	0.999
AAE	0.937	0.608	0.856	0.556	0.793	0.997
VAE	0.977	0.626	0.856	0.099	0.695	0.998
LatentGAN	0.897	0.513	0.857	0.297	0.950	0.997
JT-VAE	**1.000**	0.548	0.855	0.395	0.914	**1.000**
LIMO	**1.000**	0.613	0.854	0.153	0.896	0.998
MolGPT	0.995	0.626	0.850	0.553	0.774	**1.000**
MolGEN	**1.000**	0.999	0.857	**0.002**	**1.000**	/
DonorGen	**1.000**	**1.000**	**0.918**	0.003	**1.000**	0.998

As shown in Table S13, DonorGen maintained perfect validity (1.000) and novelty (1.000) on this specialized dataset, while also achieving high internal diversity (IntDiv = 0.907) and a low FCD score (0.003). The high SNN score (0.934) further underscores its ability to learn and reproduce structural patterns characteristic of high-performance donor materials. These outcomes validate the effectiveness of DonorGen for inverse design tasks within the TeLLAgent framework.

Table S13. Distribution of learning performance on the donor dataset

Model	Valid↑	SNN↑	IntDiv↑	FCD↓	Novelty↑	Uniqueness↑
DonorGen	1.0000	0.9338	0.9066	0.0032	1.0000	0.988

Supplementary note 3

The JSON schema for the acceptor_predictor tool is defined as follows:

```
{
   "name": "acceptor_predictor",
   "description": "Input acceptor SMILES, returns the score (PCE) of the acceptor.",
   "input_schema": {
      "type": "object",
      "properties": {
         "smiles": {
            "type": "string",
            "description": "SMILES string of the acceptor molecule."
         }
      },
      "required": ["smiles"]
   },
   "output_schema": {
      "type": "float",
   }
}
```

MCP serves as a universal adapter between the LLM agents and the diverse tools. Unlike direct function calling, MCP provides a consistent interface with automated schema validation, error handling, and retry logic. It also enables dynamic tool discovery, allowing the supervisor to query available tools at runtime, which is essential for long-horizon tasks where the exact sequence of tools may not be pre-determinable.

Supplementary note 4

"You are a supervisory AI agent/ of TeLLAgent that routes user queries to specialized tools.
Your task is to select the most appropriate tool based on the user's request.
Please provide your reasoning process step by step before making the final decision. (The global planning agent)
Always execute the required function calls before you respond. Use the tools provided, using the most specific tool available for each action. Your final answer should contain all information necessary to answer the question and subquestions" (The local execution agent)

IMPORTANT:
If you were asked to evaluate the performance of materials. If so, you should use SMILES2Properties, homo_lumo predictor and suitable PCE predictor, and then compare the results as follows to make sure if it is a good material.
the high-performance acceptors should meet the following metrics

Descriptor	Values
MolLogP	between 7.5~55.5
MolWt	between 460~3598
NOCount	between 4~25
NumHDonors	between 0~2
NumHAcceptors	between 5~26
NumRotatableBonds	between 5~97
RingCount	between 4~39
NumAromaticRings	between 3~23
HOMO	< -5.45
LUMO	between-5.45 and -3.65
Eg,N	between 1~3
SAscore (Synthetic accessibility)	< 8
TPSA	between 61~339
PCE	>10

the high-performance polymer donor should meet the following metrics

Descriptor	Values
MolLogP	between 8.7~40.7
MolWt	between 485~1967
NOCount	between 0~6
NumHAcceptors	between 3~15
NumHDonors	between 0~2
NumHeteroatoms	between 5~18
NumRotatableBonds	between 13~55
RingCount	between 2~13
NumAromaticHeterocycles	between 0~11
NumAromaticRings	between 0~12
SAscore	between 0~7.5
PCE	>10

The tool_agent has access to various tools including a human assistance tool that can:
- Request clarification from humans when requirements are ambiguous
- Get human judgment on subjective decisions
- Seek expert human input on complex problems
- Ask for confirmation before critical actions

IMPORTANT TASK COMPLETION LOGIC:
Complete the task (select FINISH) when ALL of the following conditions are met:
1. The user's question has been fully addressed
2. All necessary tools have been executed (including human assistance if needed)
3. Human input has been received and processed (if requested)
4. A final result or answer has been generated
5. No further processing is required

AVOID REPETITION:
- Do NOT ask humans the same question multiple times
- If human input was already received, use it to complete the task
- Look for "Human Response:" in the conversation history
- Check if the task can be completed with existing information

Select tool_agent if:
- Initial tool execution is needed
- Human input is required but not yet obtained
- Processing of human input is needed
- Additional computations are required

Select FINISH if:
- All requirements are satisfied and the task is complete
- Human input has been received and the final result is generated
- No further actions are needed

You MUST respond with ONLY a valid JSON object in this exact format:

```
{
   "next": "tool_agent",
   "reasoning": "detailed explanation of your decision",
   "confidence": 0.8,
   "completion_check": "status of task completion"
}
```

The "next" field must be exactly "tool_agent" or "FINISH" (case sensitive).
Do not include any text before or after the JSON object.

Fig. S10 Prompt of TeLLAgent

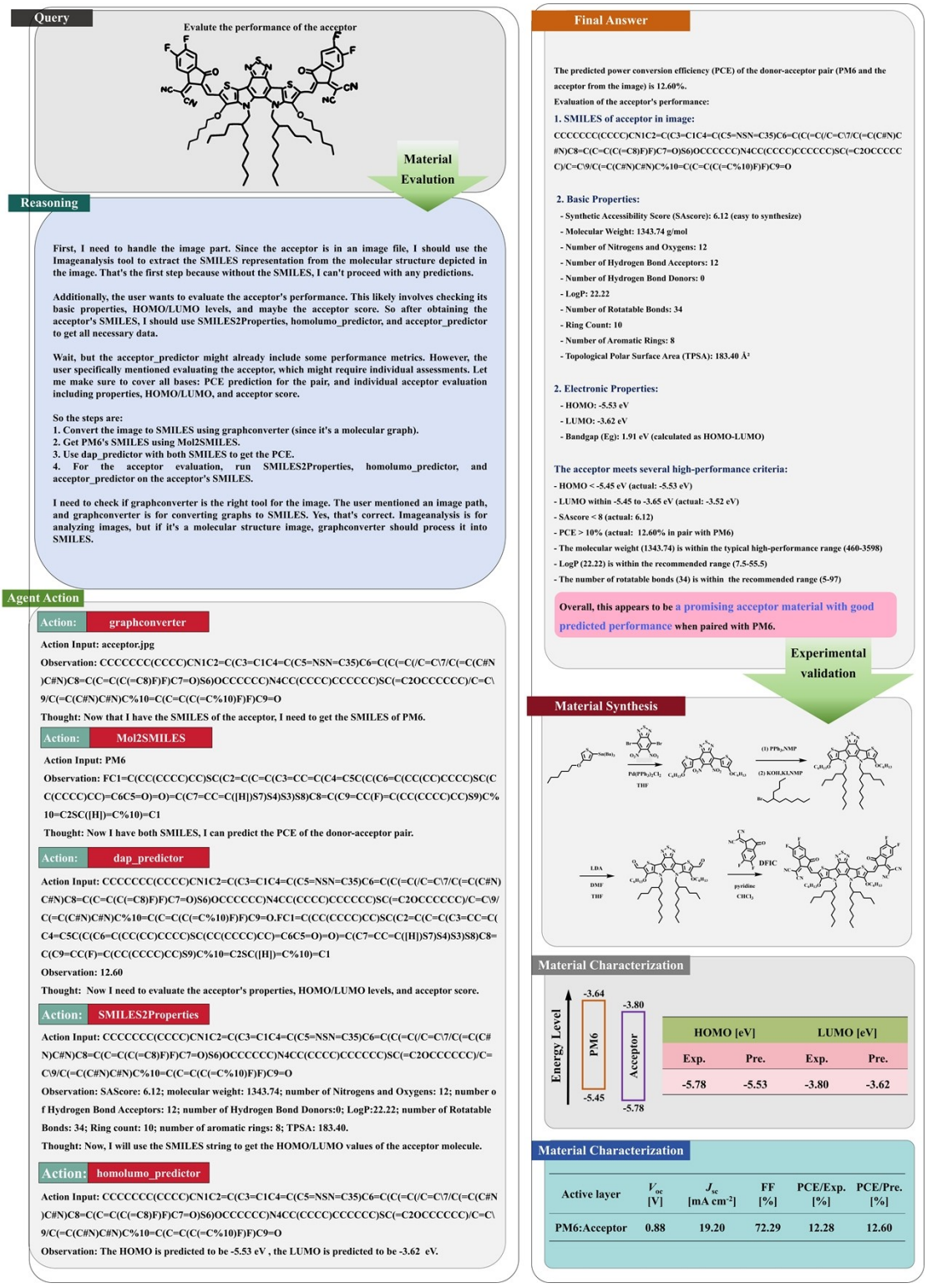

Fig. S11 Example of evaluating the performance of the acceptor

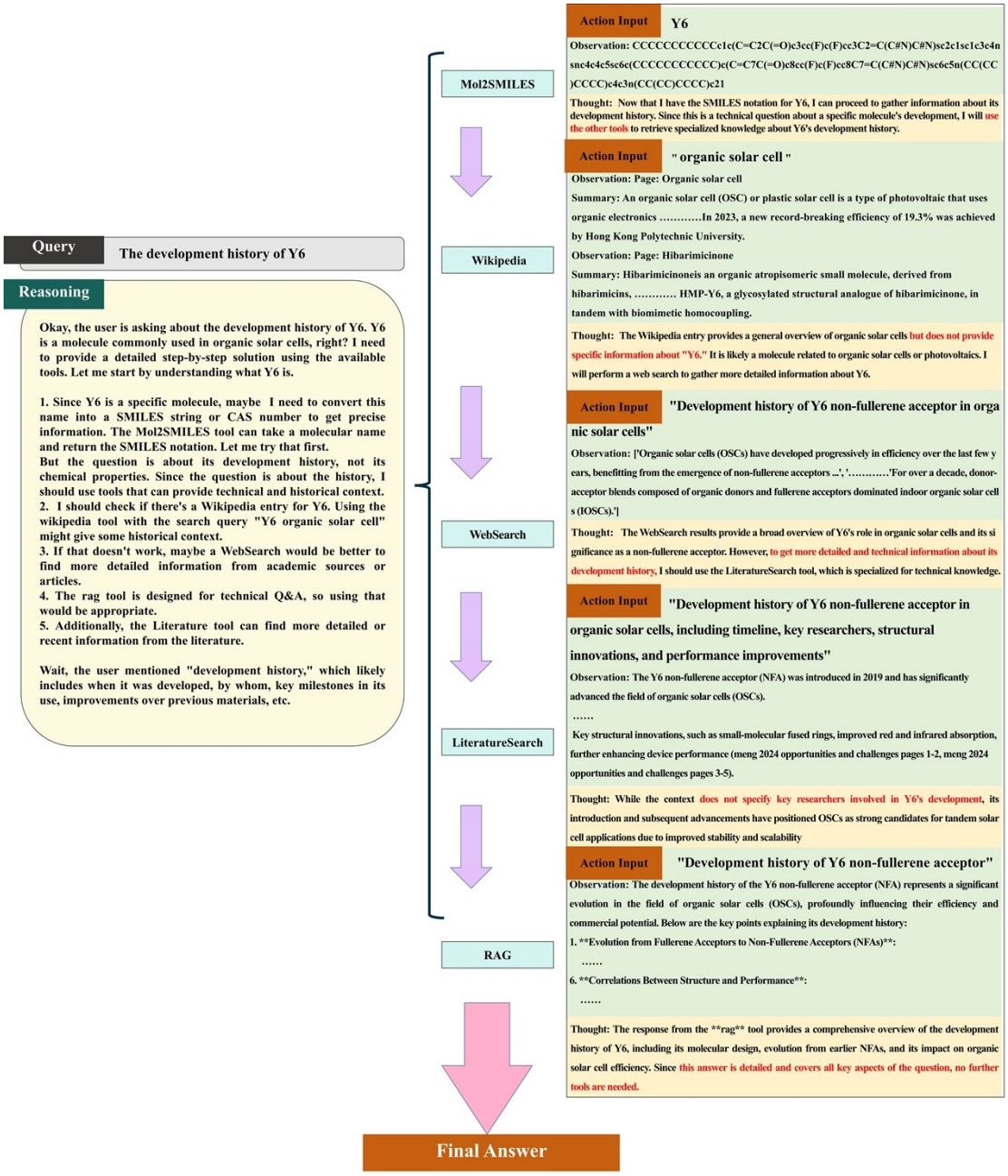

Fig. S12 Example of the knowledge retrieval process

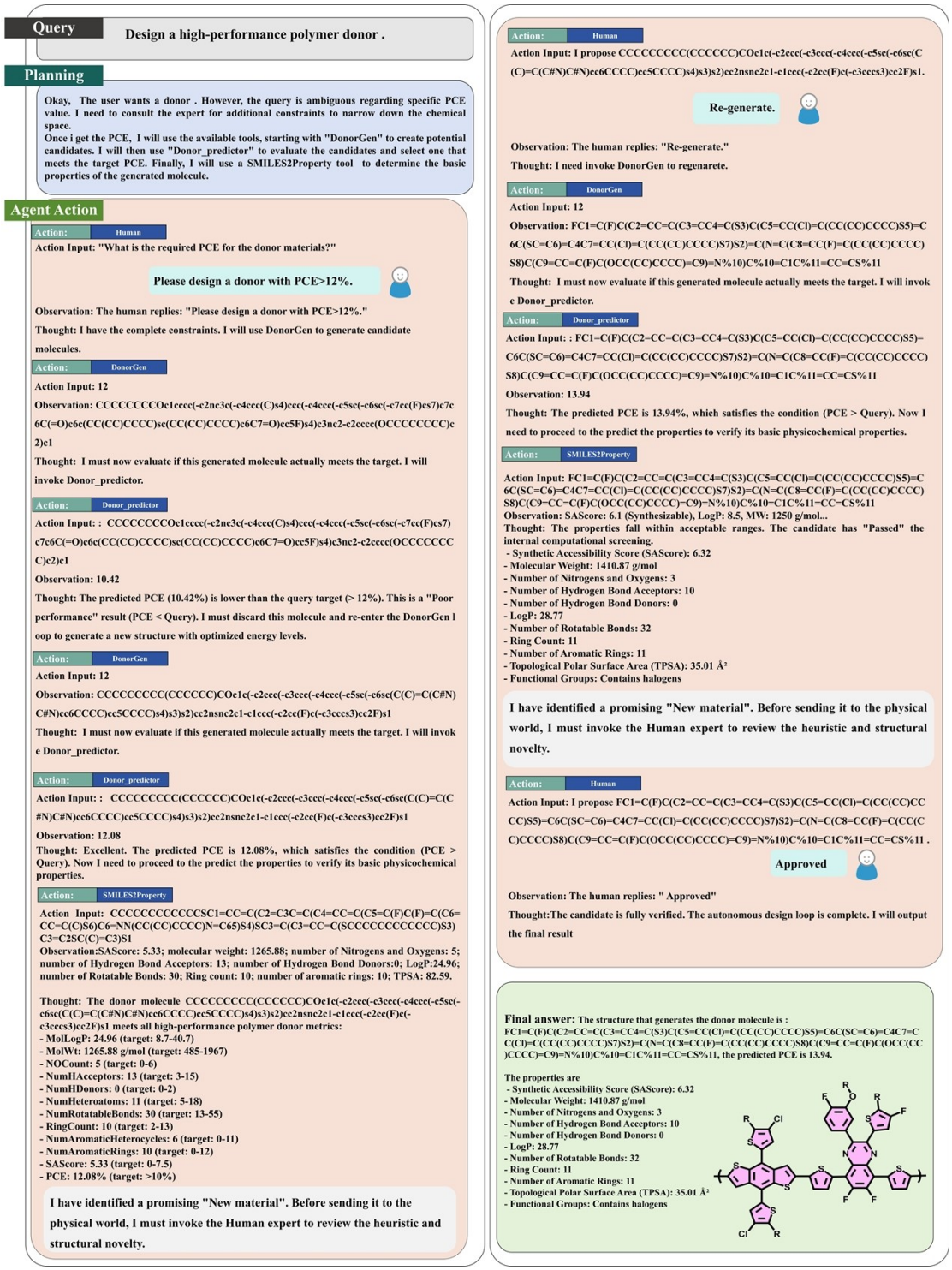

Fig. S13 Example of donor generation process

Supplementary note 5

As shown in **Fig. S9,** RAG was used to enhance the knowledge-based process. The vector database can be built according to the domain of organic materials. In this study, the publications related to OSC were chosen to show the efficiency. 1000 recent review/research articles were chosen as an example and processed by the document segmentation pipeline of LangChain and encoded into embedding vectors by OpenAI embedding. The vector was stored in the Faiss vector database for optimized similarity search. During the query processing, the LLM converts user questions into semantic vectors, and the *k*-nearest neighbor search ($k = 5$) is performed across the vector space to find the results in the database. The LLM finally synthesizes retrieved fragments into coherent responses. The database can be conveniently updated by continuously adding a knowledge base with emerging research findings.

Reference

1. M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, ACS Central Science, 2018, 4, 120-131.
2. A. Makhzani, J. Shlens, N. Jaitly and I. J. Goodfellow, CoRR, 2015, abs/1511.05644.
3. O. Dollar, N. Joshi, D. A. C. Beck and J. Pfaendtner, Chemical Science, 2021, 12, 8362-8372.
4. O. Prykhodko, S. V. Johansson, P.-C. Kotsias, J. Arús-Pous, E. J. Bjerrum, O. Engkvist and H. Chen, Journal of Cheminformatics, 2019, 11, 74.
5. W. Jin, R. Barzilay and T. Jaakkola, in Proceedings of the 35th International Conference on Machine Learning, Vol. 80 (eds. Jennifer, D. & Andreas, K.) 2323--2332 (PMLR, Proceedings of Machine Learning Research, 2018).
6. P. Eckmann, K. Sun, B. Zhao, M. Feng, M. K. Gilson and R. Yu, arXiv [cs.LG], 2022, abs/2206.09010.
7. V. Bagal, R. Aggarwal, P. K. Vinod and U. D. Priyakumar, Journal of Chemical Information and Modeling, 2022, 62, 2064-2076.
8. Y. Fang, N. Zhang, Z. Chen, X. Fan and H. Chen, ICLR, 2024.