Workflow-R1: Group Sub-sequence Policy Optimization for Multi-turn Workflow Construction

Mingze Kong [1] **Zikun Qu** [1] **Zhongquan Zhou** [2] **Pengyu Liang** [3] **Xiang Li** [3] **Zhiwei Shang** [1] **Zhi Hong** [1] **Kaiyu Huang** [4] **Zhiyong Wang** [5] **Zhongxiang Dai** [1]

Abstract

The rapid evolution of agentic workflows has demonstrated strong performance of LLM-based agents in addressing complex reasoning tasks. However, existing workflow optimization methods typically formulate workflow synthesis as a static, one-shot code-centric generation problem. This paradigm imposes excessive constraints on the model's coding capabilities and restricts the flexibility required for dynamic problem-solving. In this paper, we present Workflow-R1, a framework that reformulates workflow construction as a multi-turn, natural language-based sequential decision-making process. To resolve the optimization granularity mismatch inherent in such multi-turn interactions, we introduce Group Subsequence Policy Optimization (GSsPO). While explicitly tailored to align with the interleaved Think-Action dynamics of agentic reasoning, GSsPO fundamentally functions as a structure-aware RL algorithm generalizable to a broad class of multi-turn agentic sequential decision-making tasks. By recalibrating the optimization unit to the composite sub-sequence, specifically the atomic Think-Action cycle, it aligns gradient updates with the semantic boundaries of these interactions, ensuring robust learning in complex multi-turn reasoning tasks. Through extensive experiments on multiple QA benchmarks, Workflow-R1 outperforms competitive baselines, validating GSsPO as a generalized solution for sequential reasoning and establishing Workflow-R1 as a promising new paradigm for automated workflow optimization.

[1]The Chinese University of Hong Kong, Shenzhen [2]Shanghai Jiao Tong University [3]Tianjin University [4]Tongji University [5]University of Edinburgh. Correspondence to: Zhongxiang Dai <daizhongxiang@cuhk.edu.cn>.

Accepted at the Workshop on Reinforcement Learning from World Feedback (RLxF) at the International Conference on Machine Learning (ICML 2026), Seoul, South Korea.

1. Introduction

Large Language Models (LLMs) have evolved far beyond simple conversational interfaces. By integrating with external tools and structured reasoning paths, they act as autonomous agents capable of solving complex problems (Hong et al., 2024; Wu et al., 2024). The industrial focus has increasingly shifted towards Workflow-based Agents, defined as systems that decompose complex queries into executable graphs of operators such as search, answer, review, and revise. This paradigm has emerged as a robust standard for tackling intricate tasks.

Recognizing this potential, a growing body of research pursues automated workflow optimization, aiming to replace manual prompt engineering with algorithms that autonomously discover optimal interaction patterns. Many recent methods generate a complete workflow program (often as executable code) to orchestrate operators in a single pass, before any operator is executed and its intermediate results are observed (Zhang et al., 2025d; Hu et al., 2024; Nie et al., 2025; Zhang et al., 2025f; Wang et al., 2025a; Zhang et al., 2025a; Gao et al., 2025; Zhang et al., 2025b; Ye et al., 2025; Wang et al., 2025b; Su et al., 2025; Zhou et al., 2025). While effective for well-defined tasks, this paradigm imposes a fundamental limitation which we term the **Static Execution Trap**.

In these frameworks, the computational graph is fully determined prior to execution, effectively decoupling the planning phase from the runtime execution. This results in an open-loop system: the agent commits to a complete sequence of operators without access to intermediate observations or operator execution results generated during the process. Consequently, the workflow lacks the adaptability to handle dynamic environments, as the control flow remains rigid regardless of evolving observations. A more principled approach treats workflow construction not as static program synthesis, but as a sequential decision-making process, where the policy is continuously conditioned on the observation trajectory.

To overcome this inherent rigidity, we draw inspiration from recent advances in iterative reasoning (Li et al., 2025; Zhang

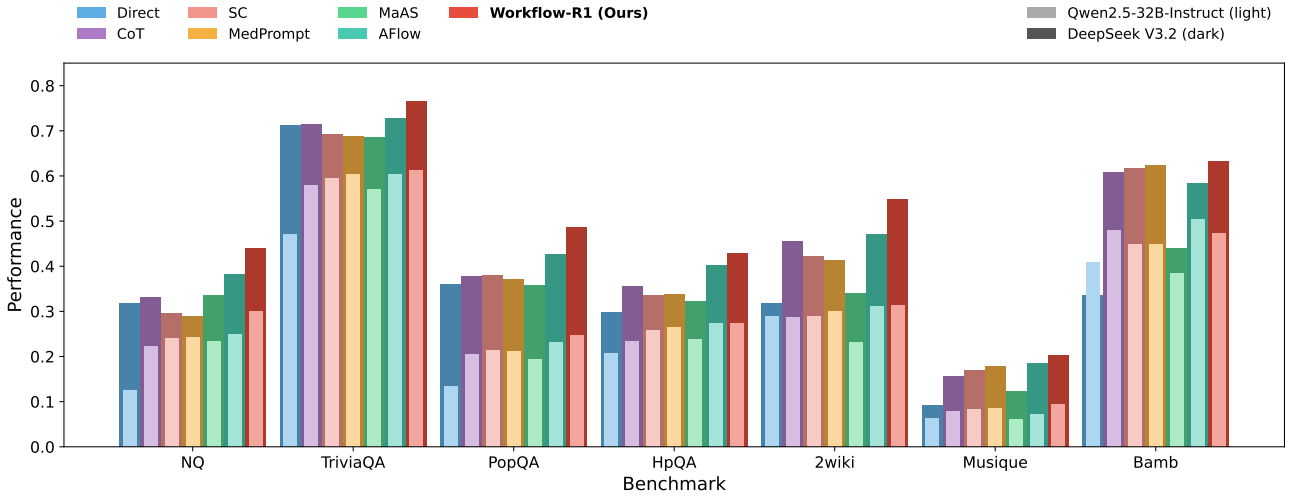

Figure 1. Performance comparison of `Workflow-R1` against SOTA baselines across seven benchmarks. Lighter and darker shades denote results using Qwen2.5-32B-Instruct and DeepSeek V3.2 as backbones, respectively. `Workflow-R1`demonstrates superior performance across the evaluated benchmarks compared to both standard prompting strategies and advanced workflow optimization methods.

et al., 2025c; Jin et al., 2025; Feng et al., 2025; Zhang et al., 2025e) and introduce `Workflow-R1` . This novel framework facilitates a paradigm shift from static planning to **Dynamic Interaction**. Specifically, we reframe workflow optimization as a multi-turn conversation between the agent and the environment. Instead of generating a monolithic script, the agent engages in an interleaved cycle of *Thinking, Acting, and Observing*. This closed-loop design ensures that each decision is grounded in operator execution results from preceding steps, empowering the model to dynamically adjust its reasoning path and operator selection in real-time.

Furthermore, in `Workflow-R1`, this interaction is conducted via natural language rather than executable code. This design yields a critical structural advantage: by relieving the agent of the strict constraints of syntactic precision and variable management inherent in code generation, the optimization task becomes significantly more tractable. This abstraction simplifies the decision process, directing the model's focus purely on logical reasoning and operator synergy. Crucially, this simplification enables a cold-start capability: the agent can learn to optimize workflows from scratch via Reinforcement Learning (RL) efficiently, bypassing the need for expert code demonstrations or Supervised Fine-Tuning (SFT) warm-ups.

However, aligning this dynamic, multi-turn decision-making process with reinforcement learning presents a unique challenge: a **Granularity Mismatch** between standard optimization objectives and the intrinsic semantic structure of agentic reasoning.

Current approaches operate at granularities that create a mismatch with the sequential nature of problem-solving, falling into two problematic extremes. On one end, token-level methods, such as GRPO (Shao et al., 2024), apply optimization at the individual token level. This approach effectively treats complex reasoning chains as loose collections of independent tokens. By enforcing updates on isolated tokens rather than coherent logical units, it risks disrupting the semantic integrity of the think-action process, while overlooking the strong causal dependencies that link a reasoning thought to its subsequent action. On the other end, sequence-level approaches, such as GSPO (Zheng et al., 2025), treat the entire multi-turn interaction sequence as a monolithic unit. While this aligns the optimization with the final reward signal, it obscures the distinct decision steps inherent in multi-turn workflows. In long-horizon tasks, relying on a single, coarse-grained signal for the whole sequence fails to differentiate effective reasoning steps from erroneous ones, resulting in inefficient learning and slow policy improvement.

We argue that the optimal alignment target is neither the single token nor the holistic sequence, but the Sub-sequence, defined as a decision unit corresponding to a contiguous segment parsed from the full response. In our multi-turn setting, each turn yields a Think-Action pair, which by construction forms a Sub-sequence. To realize this, we propose **Group Sub-sequence Policy Optimization (GSsPO)**. This structure-aware RL algorithm explicitly aligns the optimization granularity with the agent's decision unit. By treating each Sub-sequence as an atomic optimization unit, GSsPO bridges the gap between fine-grained feedback and global consistency, ensuring that policy updates strictly adhere to the decision boundaries of agentic reasoning.

In summary, this work advances the frontier of automated workflow optimization through three primary contributions:

1. A Paradigm Shift to Dynamic Interaction. We propose `Workflow-R1`, which transforms workflow construction from static program synthesis to a dynamic sequential decision process.

2. Structure-Aware Policy Optimization. We introduce **Group Sub-sequence Policy Optimization (GSsPO)** to address the granularity mismatch in multi-turn reasoning. Crucially, GSsPO serves as a generalizable algorithmic solution, capable of adapting to and enhancing the sequential reasoning capabilities of diverse multi-turn agent systems.

3. Empirical State-of-the-Art. We demonstrate that `Workflow-R1` achieves state-of-the-art performance on complex benchmarks. These results empirically validate the effectiveness of our novel dynamic framework, establishing it as a promising new paradigm for automated workflow optimization.

2. Preliminaries

In this paper, we formalize the large language model, parameterized by θ, as a policy π_θ. Given an input prompt x sampled from a dataset $\mathcal{D}$, the model generates a response $y = [y_1, y_2, \ldots, y_T]$, where y_t denotes the token at step t. The joint probability is factorized as $\pi_\theta(y|x) = \prod_{t=1}^{T} \pi_\theta(y_t|y_{<t}, x)$.

2.1. Group Relative Policy Optimization (GRPO)

GRPO obviates the need for the value model by estimating advantages through group-wise sampling (Shao et al., 2024). For each input, it generates multiple outputs and computes relative advantages by normalizing rewards within the group. The optimization objective is defined at the token-level (we omit the KL regularization for brevity):

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \min\left(r_{i,t}(\theta)\widehat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_{i,t}\right)\right], \tag{1}$$

where the token-level importance sampling ratio $r_{i,t}(\theta)$ is defined as:

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|y_{i,<t}, x)}{\pi_{\theta_{\text{old}}}(y_{i,t}|y_{i,<t}, x)}. \tag{2}$$

Consistent with prior work, all methods in this paper employ a group-based advantage estimation, where the advantage is computed by standardizing the reward within each sampled group:

$$\widehat{A}_{i,t} = \frac{r(x, y_i) - \text{mean}\left(\{r(x, y_j)\}_{j=1}^{G}\right)}{\text{std}\left(\{r(x, y_j)\}_{j=1}^{G}\right)}, \tag{3}$$

where $r(x, y_i)$ denotes the reward for the i-th sampled response conditioned on input x. Consistent with outcome supervision RL, this formulation uniformly propagates the response-level advantage to every token within the generated response.

2.2. Group Sequence Policy Optimization (GSPO).

In contrast to GRPO's token-level optimization, GSPO treats the entire response sequence as the optimization unit (Zheng et al., 2025). Accordingly, it applies the importance sampling ratio at the sequence-level, aligning gradient updates with overall generation quality:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(s_i(\theta)\widehat{A}_i, \text{clip}\left(s_i(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{A}_i\right)\right], \tag{4}$$

where the sequence-level importance sampling ratio $s_i(\theta)$ is the geometric mean of the per-token likelihood ratios for the full response:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t}|y_{i,<t}, x)}{\pi_{\theta_{\text{old}}}(y_{i,t}|y_{i,<t}, x)}\right). \tag{5}$$

The advantage $\widehat{A}_i$ follows the same group-normalized formulation as in GRPO, maintaining consistent reward normalization.

3. Workflow-R1

In this section, we present the Workflow-R1 framework. We begin with Group Sub-sequence Policy Optimization, our core RL algorithm. We then formalize the multi-turn workflow construction task in the RL setting. Next, we define the reward function for the multi-turn workflow construction. Finally, we specify the prompt template, operator pool and operator interface protocols.

3.1. GSsPO: Group Sub-sequence Policy Optimization

Motivation. Existing RL paradigms for LLMs operate at granularities that create a fundamental mismatch with the sequential nature of multi-turn decision-making. Token-level optimization, exemplified by GRPO, treats complex reasoning chains as loose collections of independent tokens. By enforcing updates on isolated tokens rather than coherent decision units, it risks disrupting the semantic integrity of the decision-making process. Conversely, sequence-level optimization, such as GSPO, treats the entire interaction as

a decision unit. While this aligns with the final reward, it obscures distinct decision boundaries, failing to differentiate effective decision units from erroneous ones.

To bridge this gap, we posit that optimization granularity must strictly align with the agent's decision boundaries. Therefore, we propose **Group Sub-sequence Policy Optimization (GSsPO)**. The core design of GSsPO recalibrates importance sampling boundaries to coincide with parsed sub-sequences, which are defined as contiguous segments corresponding to the atomic decision unit. By grounding optimization at this level, GSsPO ensures that gradient updates respect the logical coherence of each decision unit. Furthermore, to mitigate verbosity bias, we explicitly neutralize the confounding factors of action cardinality. This ensures that the optimization process is driven by the quality of the decision unit rather than its length, significantly enhancing model performance in complex multi-turn workflows.

Objective Function. Our proposed method, GSsPO, recalibrates the optimization granularity to sub-sequences to enhance policy performance. Rather than aggregating gradients at the token or sequence level, GSsPO treats each constituent sub-sequence as a discrete atomic optimization unit. Formally, for a sampled response y_i, let $\mathcal{S}_i$ denote the set of sub-sequences parsed from response i. The optimization objective function is defined as follows:

$$\mathcal{J}_{\text{GSsPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\min\left(r_s(\theta)\widehat{A}_s, \text{clip}(r_s(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_s\right)\right], \tag{6}$$

where the inner summation iterates over all sub-sequences s parsed from the sampled sequences. The advantage term $\widehat{A}_s$ inherits the sequence-level outcome, meaning all sub-sequences within the same response share the identical group-normalized advantage value $\widehat{A}_i$. The sub-sequence-level importance sampling ratio $r_s(\theta)$ is formulated as the geometric mean of the token probability ratios:

$$\begin{aligned} r_s(\theta) &= \left(\frac{\prod_{t\in s}\pi_\theta(y_t|y_{<t},x)}{\prod_{t\in s}\pi_{\theta_{\text{old}}}(y_t|y_{<t},x)}\right)^{\frac{1}{|s|}} \\ &= \exp\left(\frac{1}{|s|}\sum_{t\in s}\log\frac{\pi_\theta(y_t|y_{<t},x)}{\pi_{\theta_{\text{old}}}(y_t|y_{<t},x)}\right). \end{aligned} \tag{7}$$

Policy Gradient. To optimize the policy parameters θ, we compute the gradient of the optimization objective $\mathcal{J}_{\text{GSsPO}}$. The policy gradient is derived as follows (clipping is omitted for brevity):

$$\begin{aligned} \nabla_\theta\mathcal{J}_{\text{GSsPO}}(\theta) &= \nabla_\theta\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}r_s(\theta)\widehat{A}_s\right] \\ &= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\widehat{A}_s\cdot\nabla_\theta r_s(\theta)\right] \\ &= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}r_s(\theta)\widehat{A}_s\cdot\nabla_\theta\log r_s(\theta)\right] \\ &= \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\underbrace{\left(\frac{\pi_\theta(s|x)}{\pi_{\theta_{\text{old}}}(s|x)}\right)^{\frac{1}{|s|}}}_{r_s(\theta)}\widehat{A}_s \\ &\quad\cdot\underbrace{\frac{1}{|s|}\sum_{t\in s}\nabla_\theta\log\pi_\theta(y_t|y_{<t},x)}_{\text{Mean Token Gradient in Sub-sequence}}\Bigg]. \end{aligned} \tag{8}$$

For comparative analysis, the gradient derivations for GRPO and GSPO are provided in appendix B. Distinct from existing RL methods, our formulation introduces two fundamental shifts in optimization mechanics. First, regarding importance sampling, GSsPO assigns a unified weight $r_s(\theta)$ to all tokens within the same sub-sequence. Second, the gradient aggregation follows a hierarchical two-stage averaging mechanism: it normalizes the token length within each sub-sequence and subsequently averages by the total sub-sequence count $|\mathcal{S}_i|$.

3.2. Multi-turn Think-Action Workflow Construction

Distinct from static paradigms that synthesize fixed execution graphs, `Workflow-R1` reformulates workflow construction as a sequential decision process. The agent autonomously navigates the workflow topology through a recursive cognitive cycle to dynamically discover optimal operator combinations, ensuring the generated topology aligns with specific query requirements. The interaction is governed by a natural language interface protocol. Upon receiving a query, the agent enters a reasoning phase enclosed within `<think>` tags to formulate a strategic decision: either to conclude via `<answer>` or to invoke an operator via `<tool>`. We formally define this **Think-Action pair** generated within each turn as the **sub-sequence** in our task. Since the distinct reasoning process and the subsequent operator call jointly form a unified coherent decision unit, we establish this entity as the atomic unit of optimization, where segmentation boundaries are unambiguous (one turn constitutes exactly one sub-sequence). The abstract Operator Execution Engine intercepts commands and returns operator execution results wrapped in `<info>` tags. This feedback closes the perception-action loop, triggering the subsequent reasoning phase. By eliminating rigid structural

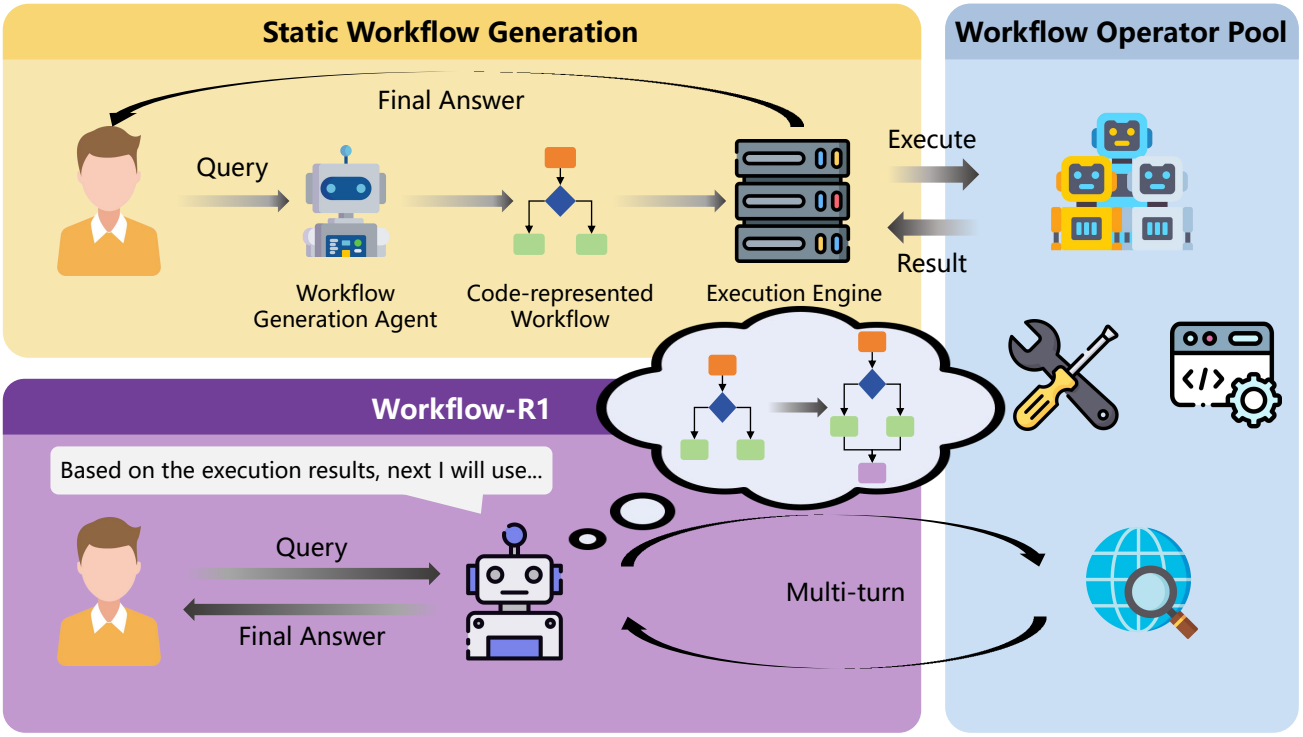

Figure 2. The top panel shows conventional static workflow generation, where the agent produces a complete executable code-represented workflow. The bottom panel depicts `Workflow-R1` multi-turn interaction, in which the agent incrementally constructs and adapts the workflow through think-action-observation cycles conditioned on execution results.

boundaries, this architecture empowers the model to aggressively explore operator synergies and adaptively orchestrate resources for deep compositional reasoning.

To enable autonomous workflow construction, we employ a reinforcement learning framework. The general optimization objective over the workflow operator pool $\mathcal{W}$ is defined as:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x;\mathcal{W})} \left[r_\phi(x,y) - \beta \log \frac{\pi_\theta(y|x;\mathcal{W})}{\pi_{\text{ref}}(y|x;\mathcal{W})} \right]. \tag{9}$$

This formulation is generic and compatible with various reinforcement learning paradigms, such as GRPO and GSPO. In this work, we adopt our proposed GSsPO as the default algorithm to optimize the workflow construction agent's policy.

3.3. Reward Design

We employ a rule-based composite reward function to jointly optimize structural validity and response accuracy. To enable the autonomous exploration of complex workflows while ensuring correctness of format, we enforce strict outcome supervision and formatting constraints.

Format Rewards. To enforce adherence to the interface protocol, we apply graded penalties $\mathcal{R}_{\text{Format}}$ based on the severity of syntactic violations:

$$\mathcal{R}_{\text{Format}} = \begin{cases} 0, & \text{if the format is completely correct} \\ -0.5, & \text{if the tag order is incorrect} \\ -1, & \text{if the <answer> tag is missing} \end{cases} \tag{10}$$

Outcome Supervision. We adopt **Exact Match (EM)** as the outcome reward to verify the correctness of the final answer against the ground truth:

$$\mathcal{R}_{\text{Outcome}}(y,g) = \text{EM}(y,g). \tag{11}$$

Consequently, the final reward signal is constructed by integrating the format constraints with the outcome supervision. The total reward $r(x,y)$ is formulated as follows:

$$r_\phi(x,y) = \mathcal{R}_{\text{Format}} + \mathcal{R}_{\text{Outcome}}. \tag{12}$$

This integration ensures that the model is guided to construct workflows within strict formatting constraints, while retaining the freedom to explore diverse operator combinations.

3.4. Operators and Prompt Template

We construct the workflow operator pool using a set of operators implemented via natural language interface definitions. This design enables the model to comprehend

Methods	Multi-turn	General QA			Multi-Hop QA				Avg.
		NQ†	TriviaQA	PopQA	HpQA†	2wiki	Musique	Bamb	
Qwen2.5-32B-Instruct									
Direct	✗	0.126	0.471	0.133	0.208	0.289	0.063	0.408	0.243
CoT	✗	0.222	0.579	0.205	0.234	0.286	0.079	0.480	0.298
SC (CoT×5)	✗	0.241	0.595	0.214	0.258	0.290	0.084	0.448	0.304
MedPrompt	✗	0.242	0.603	0.211	0.264	0.300	0.086	0.448	0.308
Agentic Workflow									
MaAS	✗	0.233	0.571	0.193	0.239	0.231	0.062	0.384	0.273
AFlow	✗	0.250	0.603	0.232	0.273	0.312	0.073	**0.504**	0.321
Workflow-R1	✓	**0.301**	**0.612**	**0.246**	**0.274**	**0.314**	**0.095**	0.472	**0.331**
Search-Augmented									
Search-R1 (GRPO)	✓	0.429	0.623	0.427	0.386	0.346	0.162	0.400	0.396
Search-R1 (PPO)	✓	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385
Workflow-R1-Search	✓	**0.435**	**0.733**	**0.493**	**0.461**	**0.633**	**0.215**	**0.576**	**0.507**

Table 1. Performance comparison of baseline methods on benchmarks. The **Multi-turn** column indicates whether the method involves iterative interaction. † represents in-domain datasets.

and execute operator calls directly through natural language instructions. Comprehensive specifications, including detailed operator usages, interface protocols, and the complete training prompt template, are provided in Appendix D.

4. Experiments

4.1. Datasets

We evaluate our `Workflow-R1` and the search-augmented version `Workflow-R1-Search` on a comprehensive suite of seven datasets designed to test the limits of agentic reasoning. The datasets are classified into two categories: (1) **General Question Answering**, including **NQ** (Kwiatkowski et al., 2019), **TriviaQA** (Joshi et al., 2017), and **PopQA** (Mallen et al., 2023); and (2) **Multi-Hop Question Answering**, comprising **HotpotQA** (Yang et al., 2018), **2WikiMultiHopQA** (Ho et al., 2020), **Musique** (Trivedi et al., 2022), and **Bamboogle** (Press et al., 2023).

Data Splits. For the training phase, we construct a dataset by sampling a subset from the training splits of NQ and HotpotQA. As detailed in Table 4, this results in a total of 10,000 training instances (4,618 from NQ and 5,382 from HotpotQA). This specific mixture is strategically designed to expose the agent to both General QA and complex multi-hop QA during the RL exploration phase.

Metric. We employ the test sets across all seven benchmarks. This comprehensive protocol allows us to assess both in-domain performance and out-of-domain generalization. We report the Exact Match (EM) score as our metric to quantify correctness.

We establish baselines on Qwen2.5-32B-Instruct (Yang et al., 2024) using **Direct Inference**, **Chain-of-Thought (CoT)** (Wei et al., 2022), **Self-Consistency (SC)** (Wang et al., 2022) with 5 paths, and **MedPrompt** (Nori et al., 2023). Regarding workflow optimization, we compare `Workflow-R1` against **AFlow** (Zhang et al., 2025d) and **MaAS** (Zhang et al., 2025a), which represent state-of-the-art task-level and query-level methods respectively. The detailed configurations for these baselines are provided in Appendix A.2. Finally, we benchmark `Workflow-R1-Search` against both PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) variants of **Search-R1** (Jin et al., 2025) to evaluate the effectiveness of our workflow optimization framework in search-augmented scenarios.

4.2. Experiments Details

Model Configuration. We employ Qwen2.5-7B-Instruct (Yang et al., 2024) as the backbone for our workflow construction agent. We utilize a fixed Qwen2.5-32B-Instruct (Yang et al., 2024) as the execution model across all agentic workflow methods, including AFlow and MaAS. Furthermore, we evaluate the performance using DeepSeek V3.2 (Liu et al., 2025) as the execution model in the ablation studies presented in Section 5.

Training Setup. We perform full-parameter fine-tuning to train two policy variants, `Workflow-R1` and `Workflow-R1-Search` , which differ strictly in their available operator pools. Unless otherwise specified, we employ GSsPO as the default RL algorithm. Table 1 reports the comprehensive benchmarks validating our framework in both standard and search-augmented scenarios, while Table 3 presents ablation studies comparing GSsPO against GRPO and GSPO. For detailed hyperparameters and configuration specifics, please refer to Appendix A.3.

Table 2. Performance comparison of methods on QA benchmarks employing **DeepSeek V3.2** as the execution model. This ablation evaluates the robustness of workflow optimization across different backbones. † represents in-domain datasets.

Methods	General QA			Multi-Hop QA				Avg.
	NQ†	TriviaQA	PopQA	HpQA†	2wiki	Musique	Bamb	
DeepSeek V3.2								
Direct	0.318	0.713	0.359	0.299	0.318	0.091	0.336	0.347
CoT	0.332	0.714	0.377	0.355	0.455	0.156	0.608	0.428
SC (CoT×5)	0.295	0.693	0.380	0.336	0.421	0.170	0.616	0.416
MedPrompt	0.288	0.688	0.370	0.338	0.413	0.179	0.624	0.414
Agentic Workflow								
MaAS	0.335	0.685	0.358	0.323	0.339	0.123	0.440	0.372
AFlow	0.382	0.727	0.426	0.402	0.470	0.185	0.584	0.454
Workflow-R1 (Ours)	0.440	**0.765**	0.487	0.428	0.548	0.202	**0.632**	0.500
Search-Augmented								
Search-R1 (GRPO)	0.429	0.623	0.427	0.386	0.346	0.162	0.400	0.396
Search-R1 (PPO)	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385
Workflow-R1-Search (Ours)	**0.456**	0.749	**0.498**	**0.460**	**0.621**	**0.211**	0.584	**0.511**

Method	NQ†	TriviaQA	HpQA†	Musique
Workflow-R1				
GRPO	0.265	0.600	0.269	0.083
GSPO	0.283	0.609	0.272	0.091
GSsPO	**0.301**	**0.612**	**0.274**	**0.095**
Workflow-R1-Search				
GRPO	0.430	0.730	0.456	0.205
GSPO	**0.457**	0.731	0.450	0.204
GSsPO	0.435	**0.733**	**0.461**	**0.215**

Table 3. Performance comparison of different RL methods across datasets.

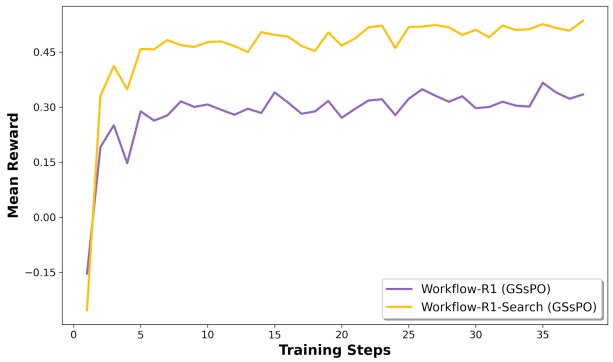

Figure 3. Mean reward convergence of GSsPO.

4.3. Experiments Analysis

Superiority in Agentic Workflow Generation. `Workflow-R1` achieves state-of-the-art performance in the standard setting, significantly outperforming both inference-time prompting baselines and static workflow optimization methods such as AFlow and MaAS. This confirms that our dynamic optimization paradigm is fundamentally more effective for reasoning tasks than static approaches. Furthermore, when equipped with the Search operator, the performance is elevated to a new level: `Workflow-R1-Search` achieves the highest scores across all baselines, establishing the SOTA on the evaluated benchmarks with an average EM of 0.507.

Case Study. Figure 4 visualizes constructed workflows that reveal the emergence of **deep reflective capabilities**, where the agent transcends rigid execution sequences by continuously monitoring operator execution results to dynamically refine its strategies. This is exemplified when the agent refrains from accepting uncertain search results at face value, instead autonomously invoking the Review operator to validate them. Moreover, upon encountering execution failures, the agent immediately adjusts its plan to attempt alternative strategies, such as leveraging PromptOptimizer to reformulate the query expression and subsequently invoking other operators to re-attempt the solution. This rational persistence enables the agent to self-correct and construct optimal solution paths for complex queries. For further demonstrations of the raw responses, please refer to Appendix E.

5. Ablation Study

Robustness with Advanced Backbone. We evaluated robustness by upgrading the backbone to DeepSeek V3.2 for both direct QA and workflow execution. This substitution yielded substantial performance gains across all agentic workflow methods. Notably, the performance advantage of automated workflow optimization over basic prompting

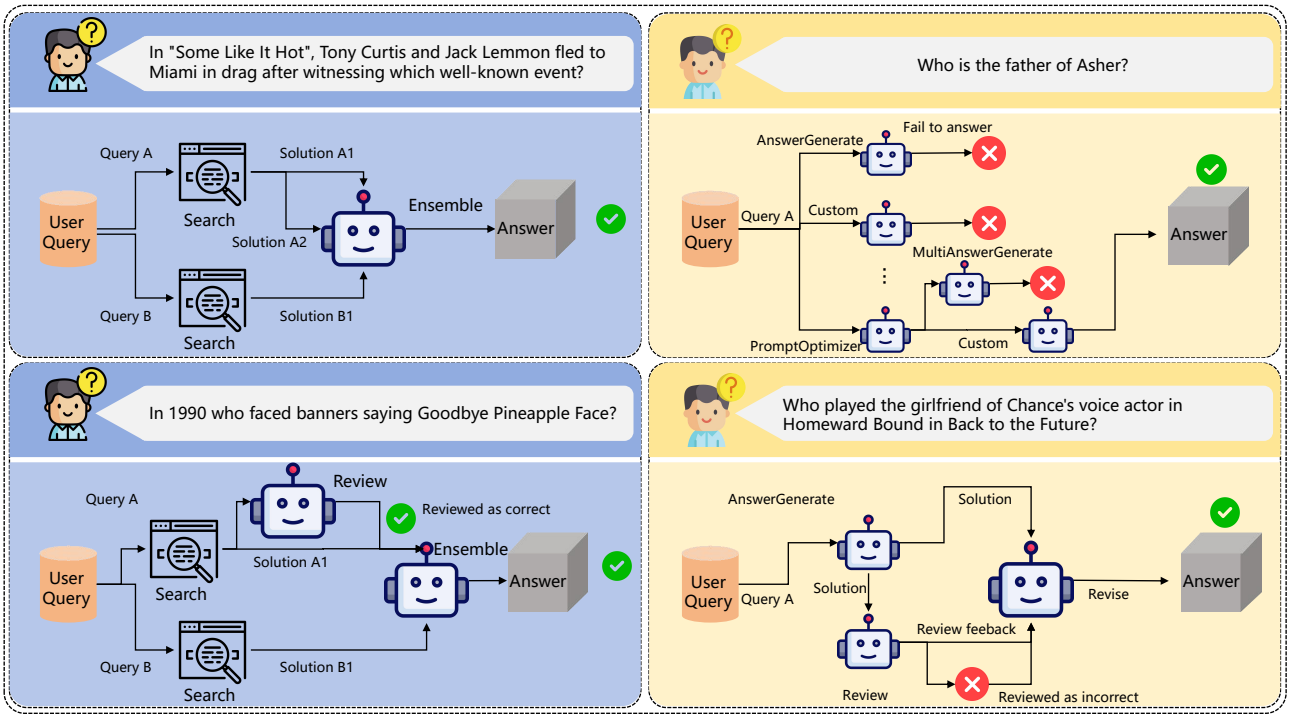

Figure 4. Visualization of workflow construction for `Workflow-R1-Search` (left, blue) and `Workflow-R1` (right, yellow).

baselines such as SC and MedPrompt increased significantly. These results indicate that superior execution models are critical for fully leveraging complex workflows to achieve qualitative performance breakthroughs.

Different Reinforcement Learning Algorithms. As shown in Table 3, GSsPO consistently outperforms GRPO across all datasets and surpasses GSPO on the majority of tasks, validating the sub-sequence as the optimal optimization unit. Furthermore, Figure 3 confirms that GSsPO maintains stable training dynamics and achieves efficient convergence.

6. Related Work

Automated Workflow Optimization. Agentic evolution spans three paradigms. Modular frameworks like DSPy (Khattab et al., 2023) and TextGrad (Yuksekgonul et al., 2024) use textual feedback, while GPTSwarm (Zhuge et al., 2024) optimizes connectivity. Code-centric ADAS (Hu et al., 2024) programs architectures via meta-agents, extended by AFlow (Zhang et al., 2025d) using MCTS. Query-adaptive methods like MAS-GPT (Ye et al., 2025), MaAS (Zhang et al., 2025a), and FlowReasoner (Gao et al., 2025) dynamically generate structures per query. However, most remain limited by a "Static Execution Trap," where pre-fixed structures cannot adapt to intermediate observations.

Reinforcement Learning for LLMs. RL aligns LLMs via methods ranging from RLHF (Ouyang et al., 2022) (e.g., PPO (Schulman et al., 2017)) to direct optimization like DPO (Rafailov et al., 2023) and RLOO (Ahmadian et al., 2024), which often face off-policy reasoning challenges (Pang et al., 2024). Recently, "Pure RL" paradigms like DeepSeek-R1 (Guo et al., 2025) and GRPO (Shao et al., 2024) eliminated critics via group rewards, with stability enhanced by DAPO (Yu et al., 2025). However, a granularity mismatch persists in multi-turn scenarios: GRPO's token-level updates risk disrupting semantic integrity, while GSPO's (Zheng et al., 2025) sequence-level signal remains too coarse for precise credit assignment.

7. Conclusion

In this paper, we present `Workflow-R1` , a framework that fundamentally transforms workflow construction from static synthesis into a dynamic sequential decision process. To enable this query-adaptive capability, we introduce Group Sub-sequence Policy Optimization (GSsPO). By establishing the Think-Action pair as the decision unit, GSsPO effectively resolves the optimization granularity mismatch inherent in conventional approaches. Extensive evaluations demonstrate that `Workflow-R1` significantly outperforms advanced baselines across seven benchmarks. Crucially, our analysis reveals the emergence of deep reflective capabilities, allowing the agent to autonomously self-correct during execution. Ultimately, Workflow-R1 establishes a robust foundation for autonomous agents, showcasing the vast potential of dynamic workflow optimization in the field of complex reasoning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Anthropic. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

Feng, J., Huang, S., Qu, X., Zhang, G., Qin, Y., Zhong, B., Jiang, C., Chi, J., and Zhong, W. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*, 2025.

Gao, H., Liu, Y., He, Y., Dou, L., Du, C., Deng, Z., Hooi, B., Lin, M., and Pang, T. Flowreasoner: Reinforcing query-level meta-agents. *arXiv preprint arXiv:2504.15257*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ho, X., Nguyen, A.-K. D., Sugawara, S., and Aizawa, A. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. Metagpt: Meta programming for a multi-agent collaborative framework. pp. et al; Google Deepmind; Google Research; Meta; Microsoft –, Hybrid, Vienna, Austria, 2024. Agent based;Collaborative framework;Complex task;Language model;Meta Programming;Model-based OPC;Multi agent;Problem-solving;Society of agents;Work-flows;.

Hu, S., Lu, C., and Clune, J. Automated design of agentic systems. *arXiv preprint arXiv:2408.08435*, 2024.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., and Han, J. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., and Dou, Z. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Nie, F., Feng, L., Ye, H., Liang, W., Lu, P., Yao, H., Alahi, A., and Zou, J. Weak-for-strong: Training weak meta-agent to harness strong executors. *arXiv preprint arXiv:2504.04785*, 2025.

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pang, R. Y., Yuan, W., He, H., Cho, K., Sukhbaatar, S., and Weston, J. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37: 116617–116637, 2024.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Su, J., Xia, Y., Lan, Q., Song, X., Jingsong, Y., He, L., and Shi, T. Difficulty-aware agent orchestration in llm-powered workflows. *arXiv e-prints*, pp. arXiv–2509, 2025.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, Y., Liu, S., Fang, J., and Meng, Z. Evoagentx: An automated framework for evolving agentic workflows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 643–655, 2025a.

Wang, Y., Yang, L., Li, G., Wang, M., and Aragam, B. Scoreflow: Mastering llm agent workflows via score-based preference optimization. *arXiv preprint arXiv:2502.04306*, 2025b.

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. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. Common sense;Expert knowledge;High quality;Knowledge capabilities;Language model;Multi-stages;Pre-training;Reasoning capabilities;Reinforcement learnings;Training dataset;, 2024. URL http://dx.doi.org/10.48550/arXiv.2412.15115.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Ye, R., Tang, S., Ge, R., Du, Y., Yin, Z., Chen, S., and Shao, J. Mas-gpt: Training llms to build llm-based multi-agent systems. volume 267, pp. 72063 – 72090, Vancouver, BC, Canada, 2025. Adaptive multiagent systems;Data construction;Executable codes;High quality;Multiagent systems (MASs);Novel task;Open-source;Reframing;System methods;User query;.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., and Zou, J. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Zhang, G., Niu, L., Fang, J., Wang, K., Bai, L., and Wang, X. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025a.

Zhang, G., Yue, Y., Sun, X., Wan, G., Yu, M., Fang, J., Wang, K., Chen, T., and Cheng, D. G-designer: Architecting multi-agent communication topologies via graph neural networks. volume 267,

pp. 76678 – 76692, Vancouver, BC, Canada, 2025b. Best choice;Collective intelligences;Communication topologies;Graph neural networks;Individual agent;Inter-agent communications;Language model;Model-based OPC;Multi-agent communications;Specific tasks;.

Zhang, H., Feng, T., and You, J. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Zhang, J., Xiang, J., Yu, Z., Teng, F., Chen, X.-H., Chen, J., Zhuge, M., Cheng, X., Hong, S., Wang, J., Zheng, B., Liu, B., Luo, Y., and Wu, C. Aflow: Automating agentic workflow generation. pp. 100351 – 100388, Singapore, Singapore, 2025d.

Zhang, S., Dong, Y., Zhang, J., Kautz, J., Catanzaro, B., Tao, A., Wu, Q., Yu, Z., and Liu, G. Nemotron-research-tool-n1: Exploring tool-using language models with reinforced reasoning. *arXiv preprint arXiv:2505.00024*, 2025e.

Zhang, Y., Liu, X., and Xiao, C. Metaagent: Automatically constructing multi-agent systems based on finite state machines. volume 267, pp. 75667 – 75694, Vancouver, BC, Canada, 2025f. 'current;Automated design methods;Communication structures;Finite states machine;Language model;Meta-agents;Multiagent framework;Multiagent systems (MASs);Tool integration;Training data;.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Zhou, H., Wan, X., Sun, R., Palangi, H., Iqbal, S., Vulić, I., Korhonen, A., and Arık, S. Ö. Multi-agent design: Optimizing agents with better prompts and topologies. *arXiv preprint arXiv:2502.02533*, 2025.

Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., and Schmidhuber, J. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

A. Experimental Setups

A.1. Datasets

We construct a mixed training set of 10,000 samples from NQ (4,618) and HotpotQA (5,382) to balance single-hop and multi-hop reasoning capabilities. For evaluation, we use seven datasets across two domains: General QA (NQ, TriviaQA, PopQA) and Multi-Hop QA (HotpotQA, 2WikiMultiHopQA, Musique, Bamboogle). The detailed statistics are shown in Table 4.

Category	Dataset	Train	Test
General QA	NQ	4,618	3,610
	TriviaQA	–	11,313
	PopQA	–	14,267
Multi-Hop QA	HotpotQA	5,382	7,405
	2WikiMultiHopQA	–	12,576
	Musique	–	2,417
	Bamboogle	–	125
Total		**10,000**	**51,713**

Table 4. Dataset Statistics.

A.2. Workflow Baselines

For the agentic workflow baselines, MaAS and AFlow represent state-of-the-art query-level and task-level workflow optimization methods, respectively. In our reproduction, we follow their official implementations: MaAS employs GPT-4o-mini (Hurst et al., 2024) and AFlow employs Claude-3.5-Sonnet (Anthropic, 2024) as their respective workflow optimization models. To ensure fair comparison, we use Qwen2.5-32B-Instruct as the unified workflow execution model for both training and testing in main experiments.

A.3. Training Configuration

We utilize **Qwen2.5-7B-Instruct** as the policy model (workflow optimization model) and **Qwen2.5-32B-Instruct** as the workflow execution model for `Workflow-R1` . All experiments perform full-parameter fine-tuning with the GSsPO algorithm using the VeRL framework (Sheng et al., 2025) on 2 NVIDIA H20 GPUs. The detailed hyperparameters are listed in Table 5.

Hyperparameter	Value
Training Epochs	2
Training Samples	10,000
Global Batch Size	512
Rollout	8
Total Training Steps	38
Learning Rate	1e-6
KL Coefficient	0.001
Clip ratio	0.2

Table 5. Training Hyperparameters for `Workflow-R1` and `Workflow-R1-Search` .

Regarding the policy variants, `Workflow-R1-Search` replaces the `MultiAnswerGenerate` operator with the `Search` operator to enable external information retrieval, while `Workflow-R1` retains the standard operator pool for self-contained reasoning.

B. Policy Gradient of More RL

In this section, we provide the detailed gradient derivations for GRPO and GSPO to facilitate the comparative analysis.

B.1. GRPO

Group Relative Policy Optimization (GRPO) operates at the **token-level**.

Optimization Objective: The objective function is defined as follows (omitting the clipping mechanism for brevity):

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} w_{i,t}(\theta)\hat{A}_{i,t} \right] \tag{13}$$

Importance Ratio Definition: The token-level importance ratio $w_{i,t}(\theta)$ is defined as:

$$w_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})} \tag{14}$$

Gradient Derivation: The policy gradient is derived by differentiating the objective function:

$$\begin{aligned}
\nabla_\theta \mathcal{J}_{GRPO}(\theta) &= \nabla_\theta \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})}\hat{A}_{i,t} \right] \\
&= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \hat{A}_{i,t}\nabla_\theta \left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})} \right) \right] \\
&= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})}\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t}) \right]
\end{aligned} \tag{15}$$

This derivation confirms that GRPO updates the policy based on individual token likelihoods, weighted by their specific token-level importance ratios.

B.2. GSPO

Group Sequence Policy Optimization (GSPO) treats the **entire response** as a single unit.

Optimization Objective: The objective function is formulated as follows (omitting the clipping mechanism for brevity):

$$\mathcal{J}_{GSPO}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}} \left[\frac{1}{G}\sum_{i=1}^{G} s_i(\theta)\hat{A}_i \right] \tag{16}$$

Importance Ratio Definition: The sequence-level importance sampling ratio $s_i(\theta)$ is defined based on the geometric mean of likelihood ratios:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)} \right)^{\frac{1}{|y_i|}} \tag{17}$$

Gradient Derivation: The gradient derivation proceeds as follows:

$$\begin{aligned}
\nabla_\theta \mathcal{J}_{GSPO}(\theta) &= \nabla_\theta \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\hat{A}_i\right] \\
&= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\nabla_\theta\left(\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\hat{A}_i\nabla_\theta\log\left(\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{old}}(y_i|x)}\right)^{\frac{1}{|y_i|}}\hat{A}_i\nabla_\theta\log\pi_\theta(y_{i,t}|x,y_{i,<t})\right]
\end{aligned} \tag{18}$$

In contrast to GRPO, GSPO applies a unified sequence-level weight to all tokens within a response, ensuring that the gradient updates are consistent with the semantic integrity of the sequence.

C. More Results

In addition to the main results using the 32B model, we also evaluate the performance of a smaller backbone, **Qwen2.5-7B-Instruct**, across the benchmarks. Table 6 presents the performance on four baseline methods.

Methods	Multi-turn	General QA			Multi-Hop QA				Avg.
		NQ†	TriviaQA	PopQA	HpQA†	2wiki	Musique	Bamb	
Qwen2.5-7B-Instruct									
Direct	✗	0.030	0.051	0.056	0.098	0.184	0.011	0.048	0.068
CoT	✗	0.122	0.102	0.121	0.173	0.202	0.043	0.312	0.154
SC (CoT×5)	✗	0.142	0.110	0.130	0.188	0.216	0.050	0.336	0.167
MedPrompt	✗	0.114	0.104	0.119	0.146	0.155	0.046	0.320	0.143

Table 6. Performance comparison of baseline methods on benchmarks. The **Multi-turn** column indicates whether the method involves iterative interaction. † represents in-domain datasets.

D. Details of Operator Description and Prompt Template

D.1. prompt template

We present the training prompt template, which specifies multi-turn protocols and strict formatting constraints, alongside the complete operator pool and invocation syntax. Notably, we introduce the WarningOp as a safeguard against formatting violations (e.g., incorrect Think-Action structure). This operator intercepts invalid outputs and returns a warning message as observation feedback, explicitly guiding the agent to self-correct in subsequent steps.

Query:Prompt Template

Answer the given question. Every time you get new information from an operator, you must conduct reasoning inside <think> and </think> before taking any further action. This means you should have multiple thinking sessions - one after each time you receive information from operators, not just one initial thinking session.

After reasoning, if you find you can utilize an operator to help you increase the likelihood of obtaining a correct answer, you can call an operator according to the operator description below. And when you call the operator, you MUST follow the corresponding format.

Available Operators: {AVAILABLE_OPERATOR_POOL}

Before each operator call, you must explicitly reason inside <think> and </think> about why you're choosing this operator, how you'll use it, and what you expect to learn. When you call an operator, the response will be returned between <info> and </info>. After receiving this response, you must again think inside <think> and </think> about what the information means, how it helps answer the original question, and what your next step should be.

CRITICAL INSTRUCTION: Do not settle for simple, single-step operator calls. You are REQUIRED to aggressively explore diverse, complex, and unconventional combinations of operators. Your goal is not just to answer, but to exhaustively explore how different operators can synergize. You must push the boundaries of what these operators can achieve together. Passive, lazy, or simplistic strategies are unacceptable; you must demonstrate deep reasoning through rich, multi-step operator orchestration to uncover their unique capabilities and "magical" combined effects.

IMPORTANT CONSTRAINT: You have a MAXIMUM of 20 operator calls. Plan your strategy wisely - you must gather sufficient information and reach a confident answer within this limit.

Always follow this cycle: think, then optionally use an operator if needed, then when you receive information, think again, then decide your next step. Repeat this cycle until you have enough information to answer confidently. After you have gathered enough information and conducted deep reasoning through multiple thinking cycles, if you are truly certain you can answer the query correctly, directly provide your final answer in only a few words or a short phrase inside <answer> and </answer> to exactly answer the query. If you're not certain, continue gathering information and reasoning until you reach certainty.

Query: {USER_QUERY}

Operators Description and Usage

AnswerGenerate:
Description: Call an advanced LLM to think and answer questions. Useful for domain-specific questions, knowledge QA, reasoning tasks, etc.
Usage: <tool>AnswerGenerate: YOUR_QUESTION</tool>

Custom:
Description: Call an LLM to execute any task you define. As long as you provide a clear and detailed instruction along with the task content, the LLM can accomplish virtually anything you need.
Usage: <tool>Custom: <task>YOUR_TASK_CONTENT</task> <instruction>YOUR_INSTRUCTION </instruction> </tool>

PromptOptimizer:
Description: Call an LLM to refine and optimize a query's clarification by adding details and creating a step-by-step plan. Useful when the original query is ambiguous or needs more context.
Usage: <tool>PromptOptimizer: YOUR_QUERY</tool>

Review:
Description: Call an LLM to critically review a solution's correctness and provide feedback. Useful when you want to verify whether some answer is correct before finalizing.
Usage: <tool>Review: <question>ORIGINAL_QUESTION</question><solution>SOLUTION_TO_REVIEW </solution> </tool>

Revise:
Description: Call an LLM to revise a solution based on feedback. Useful when you have received critical feedback indicating there are something wrong in the solution.
Usage: <tool>Revise: <question>ORIGINAL_QUESTION</question><solution>SOLUTION_TO_REVISE </solution> <feedback>FEEDBACK_RECEIVED</feedback> </tool>

MultiAnswerGenerate:
Description: Call an advanced LLM to generate multiple independent answers (3 times) for the same question. Useful when you want to explore diverse reasoning paths or prepare candidates for ensemble selection.
Usage: <tool>MultiAnswerGenerate: YOUR_QUESTION</tool>

ScEnsemble:
Description: Call an LLM to select the best answer from multiple candidate solutions through self-consistency ensemble selection. You must provide at least two candidate solutions.
Usage: <tool>ScEnsemble: <query>ORIGINAL_QUESTION</query><solution>CANDIDATE_1</solution> ...<solution>CANDIDATE_N</solution> </tool>

Search:
Description: Call a search engine to search for information on the web. Useful when you need to look up facts that are not already available to you.
Usage: <tool>Search: YOUR_QUERY</tool>

WarningOp:
Description: A special operator that returns a warning message when the model fails to follow the required output format.
Output: WARNING!!! I did not follow the required format. If I want to use an operator, I should use <tool>OperatorName: input</tool>. If I want to give the final answer, I should use <answer>...</answer>.

E. More Case Studies

In this section, we present a comprehensive set of qualitative examples, comprising four cases utilizing `Workflow-R1-Search` and four cases utilizing `Workflow-R1` . These case studies underscore several compelling emergent behaviors inherent to our framework. We observe that when an operator fails to return valid information due to internal errors, the agent avoids stagnation, instead intelligently pivoting to alternative strategies to resolve the query. Furthermore, we identify robust self-correction mechanisms; upon triggering format warnings, the agent explicitly reflects on the error within its reasoning traces (inside `<think>` tags). Collectively, these emergent behaviors attest to the practical value of our proposed paradigm, marking a bold exploration into the frontier of autonomous multi-turn workflow construction.

Query: **What would Kevin Scale have been called if it had adopted the originator's first name?**

Ground Truth: **William Scale**

[ASSISTANT]
<think>First, I need to understand who Kevin Scale is and what the originator's first name would be in this context. However, the query does not provide enough context about Kevin Scale or the originator. I will start by searching for information about Kevin Scale to understand its context better.</think>
<tool>Search: What is Kevin Scale and what does it refer to?</tool>

[OPERATOR]
<info>Search results for 'What is Kevin Scale and what does it refer to?':
Title: Kelvin
Snippet: The Kelvin scale is an absolute temperature scale that starts at the lowest possible temperature (absolute zero), taken to be 0 K. By definition, the Celsius ...

Title: Kelvin Temperature Scale Definition
Snippet: The Kelvin temperature scale is used by scientists because they wanted a temperature scale where zero reflects the complete absence of thermal energy.

Title: Kelvin (K) — Definition & Facts
Snippet: The ... ... by shifting the Celsius scale by $-273.15°$ so that absolute zero coincides with 0 K) ...

Title: Kelvin Scale - an overview — ScienceDirect Topics
Snippet: The Kelvin scale is defined as a temperature scale that extends the degree Celsius scale down to absolute zero (0 K), representing a complete absence of ...

Title: What is the Kelvin Temperature Scale?
Snippet: Outlining what the Kelvin temperature scale is and how it is used in chemistry. Kelvin and degrees celsius are compared and absolute zero is ...</info>

[ASSISTANT]
<think>The search results do not provide any direct information about Kevin Scale. However, the results are mostly about the Kelvin temperature scale, which might be a misinterpretation of the query. I need to clarify the context of 'Kevin Scale' and the originator's first name. I will use the PromptOptimizer to refine the query for better understanding.</think>
<tool>PromptOptimizer: What would Kevin Scale have been called if it had adopted the originator's first name?</tool>

[OPERATOR]
<info>PromptOptimizer thought process: The query is asking about a hypothetical scenario where a concept or theory named K̈evin Scaleḧas its name changed based on the originator's first name. However, there seems to be a misunderstanding in the query as K̈evin Scaledöes not appear to be a known scientific or theoretical concept. It might be a fictional or misnamed concept. To refine this query, we need to clarify what K̈evin Scaleïrefers to and who the originator is. If we assume that K̈evin Scaleïs a plac... ...'s first name instead. For example, if the concept was originally named after a person with the full name J̈ohn Kevin,änd it was currently called K̈evin Scale,ẗhe query wants to know what it would be called if it were named after John instead.
PromptOptimizer answer: Assuming K̈evin Scaleïs a placeholder for a real concept named after an individual, the refined query would be: If the concept currently known as K̈evin Scaleẅas instead named after the originator's first name, what would it be called?</info>

[ASSISTANT]
<think>The PromptOptimizer provided a more specific and clear version of the query. However, it still does not give a clear answer about what K̈evin Scaleïrefers to. I need to search for a more specific context of K̈evin Scaleïto understand its origin and the originator's first name. I will now search for K̈evin Scaleïto find out what it is and who the originator is.</think>
<tool>Search: What is Kevin Scale and who is the originator?</tool>

[OPERATOR]
<info>Search results for 'What is Kevin Scale and who is the originator?':
Title: Kelvin
Snippet: The 19th-century British scientist Lord Kelvin first developed and proposed the scale. It was often called the äbsolute Celsiusscale in the early 20th ...

Title: Celsius and Kelvin scales — Research Starters
Snippet: The Kelvin scale is another common temperature scale. British inventor and scientist William Thomson, also known as Lord Kelvin, developed the scale in 1848.

Title: Thomson, William
(Lord Kelvin... ... but is best remembered as the originator of the kelvin* temperature scale. In 1848 Thomson proposed a ...

Title: Lord Kelvin

Snippet: ... inventor Wilhelm Röntgen. ... Instead, he was named Lord Kelvin, a name that now lives on as the Kelvin scale for absolute temperature, an ...

Title: How was Kelvin discovered and named after Lord Kelvin?
Snippet: In 1848, British mathematician and scientist William Thomson (also known as Lord Kelvin) proposed an absolute temperature scale where the lowest possible ...</info>

[ASSISTANT]
<think>The search results indicate that K̈elvinïs the scale named after Lord Kelvin, a 19th-century British scientist. The query seems to be asking about a hypothetical scenario where the scale is named after the originator's first name instead of the full name K̈elvin.S̈ince the originator is Lord Kelvin, and his first name is William, the scale would have been called the Ẅilliam Scaleïf it had adopted the originator's first name. Therefore, I can now answer the query.</think>
<answer>William Scale</answer>

Query: `In 1990 who faced banners saying Goodbye Pineapple Face?`

Ground Truth: `["Manuel Antonio Noriega Moreno", "Manuel Noreiga", ...]`

[ASSISTANT]
<think>First, I need to find out who faced banners saying "Goodbye Pineapple Face" in 1990. I will start with a search query.</think>
<tool>Search: In 1990 who faced banners saying Goodbye Pineapple Face</tool>

[OPERATOR]
<info>Search results for 'In 1990 who faced banners saying Goodbye Pineapple Face':
Title: World Leaders Trivia Quiz
Snippet: In 1990 who faced banners saying Goodbye Pineapple Face? remove_red_eye Show Answer. 20. Who was the youngest US Vice President of the 20th century ...

Title: Manuel Noriega
Snippet: Manuel Antonio Noriega Moreno was a Panamanian military officer and politician who was the de facto ruler of Panama from 1983 to 1989.

Title: Harvesting Prosperity: Technology and Productivity Growth
Snippet: ... ...

Title: Sustainable diets and biodiversity
Snippet: 1990 and 2015. In 2011, some countries remain far from reaching this target, and ensuring global food security persists as one of the greatest challenges of ...

Title: The Adaptability of Marketing Systems to Interventions in ...
Snippet: by MH Hounhouigan · 2014 · Cited by 38 — When pineapple production intensified in the late 1990s in an effort to stimulate economic growth, the marketing sys- tem failed to sell the increasing ...</info>

[ASSISTANT]
<think>The search results are not very clear and direct. However, one snippet from "World Leaders Trivia Quiz" seems relevant as it directly asks about banners saying "Goodbye Pineapple Face" in 1990. I need to click on the link to get the answer. But since I can't click, I will assume the answer is provided in the quiz and it's likely a trivia question about a political figure or event. I will use the Review operator to verify if the context of the trivia question can be correctly interpreted.</think>
<tool>Review: <question>In 1990 who faced banners saying Goodbye Pineapple Face?<question>

<solution>Manuel Noriega</solution></tool>

[OPERATOR]
<info>Review feedback: The provided solution, "Manuel Noriega," does not seem to be correct based on the context of the question. The term "Pineapple Face" is not commonly associated with Manuel Noriega in historical records or media coverage related to him.
A more likely scenario involves former British Prime Minister Margaret Thatcher, who was sometimes referred to as "Dagwood Bumface" or faced similar derogatory banners during her tenure, but not specifically "Goodbye Pineapple Face." It seems there might be a mix-up or misinterpretation of the event or person involved.
To improve the solution, it would be beneficial to research the specific event where someone faced banners with this exact phrase. This could involve looking into political figures from the early 1990s who were subjects of public protest or ridicule.</info>

[ASSISTANT]
<think>The review feedback suggests that my initial guess of Manuel Noriega was incorrect. The term "Pineapple Face" is not commonly associated with him. The feedback hints that the event might be related to a political figure from the early 1990s who faced such banners. Given the context, Margaret Thatcher seems a plausible candidate as she was a prominent political figure during that time and faced various forms of public protest. I will now use the Search operator to look for more specific information about banners saying "Goodbye Pineapple Face" in the context of Margaret Thatcher or similar political figures from the early 1990s.</think>
<tool>Search: Who faced banners saying Goodbye Pineapple Face in the early 1990s</tool>

[OPERATOR]
<info>Search results for 'Who faced banners saying Goodbye Pineapple Face in the early 1990s':
Title: Manuel Noriega
Snippet: Manuel Antonio Noriega Moreno was a Panamanian military officer and politician who was the de facto ruler of Panama from 1983 to 1989.

Title: Manuel Noriega. They both got pothole pineapple face.
Snippet: Mohamed looks just like this man - Manuel Noriega. They both got pothole pineapple face. · 15y · Public · Judyth E. Beerup and Rachel Kovach.

Title: Manuel Noriega, Dictator Ousted by ... ...tives protesting the use of his likeness and the "Pineapple Face" moniker in a "Call of Duty" video game.) Embracing his ...

Title: General Noriega sues Call Of Duty for portraying him as a ' ...
Snippet: Former Panama dictator General Noriega sues Call Of Duty makers for portraying him as a 'kidnapper, murderer and old pineapple face'.

Title: Dirty Manuel
Snippet: The story of General Noriega's rise and fall inspired several good nonfiction books in the early 1990's, and it clearly delighted Lawrence ...</info>

[ASSISTANT]
<think>The search results consistently point towards Manuel Noriega and his association with the term "Pineapple Face." However, the specific banners saying "Goodbye Pineapple Face" seem to be a misunderstanding or misattribution. The feedback and search results suggest that the term "Pineapple Face" was used in a different context, possibly related to Margaret Thatcher or another political figure. Given the persistent association of "Pineapple Face" with Manuel Noriega and the lack of direct evidence for the specific banners in question, I need to re-evaluate the context and consider the possibility that the question might be referring to a different event or figure. Since the term "Pineapple Face" is strongly linked to Manuel Noriega, and the banners saying "Goodbye Pineapple Face" might be a misinterpretation, I will consider the context of the question and the feedback. The feedback suggests

that the term ”Pineapple Face” was used in a different context, possibly related to Margaret Thatcher. However, the direct association with banners saying ”Goodbye Pineapple Face” is not clear. I will use the ScEnsemble operator to compare the two candidate solutions: Manuel Noriega and Margaret Thatcher, to see if the collective wisdom can help me decide.</think>
<tool>ScEnsemble: <query> In 1990 who faced banners saying Goodbye Pineapple Face?<query> <solution>Manuel Noriega</solution><solution>Margaret Thatcher</solution></tool>

[OPERATOR]
<info>ScEnsemble thought process: To analyze the provided options, we need to consider the historical context of the phrase ”Goodbye Pineapple Face” and who it might refer to.
Option A: Manuel Noriega - Manuel Noriega was the de facto leader of Panama from 1983 until his arrest in 1989. He was later extradited to France and then to the United States to face charges. The term ”Pineapple Face” could be a derogatory nickname given to him, possibly due to his appearance or as part of political satire or media portra... ... being the correct answer based on the information given. However, if we must choose between the two based on the likelihood of facing such banners due to their prominence in international affairs during that period, Manuel Noriega might be slightly more plausible due to the intense media scrutiny and political opposition he faced around the time of his extradition.
However, it’s important to note that without concrete historical evidence, both options remain speculative.
ScEnsemble answer: Manuel Noriega</info>

[ASSISTANT]
<think>The ScEnsemble operator has provided a collective wisdom that Manuel Noriega is the more plausible answer based on the historical context and the intense media scrutiny and political opposition he faced around the time of his extradition. While the term ”Pineapple Face” is strongly associated with Manuel Noriega, the specific banners saying ”Goodbye Pineapple Face” might be a misinterpretation or a different event. However, given the context and the feedback, Manuel Noriega is the most likely answer. I will now provide the answer based on the collective wisdom.</think>
<answer>Manuel Noriega</answer>

Query: `Which film has the director died earlier, Miracle Of Marcelino or The Accompanist?`

Ground Truth: `["Miracle of Marcelino", "Miracle Of Marcelino"]`

[ASSISTANT]
<think>First, I need to find out when the director of ”Miracle Of Marcelino” died and when the director of ”The Accompanist” died. I will start by searching for the director of ”Miracle Of Marcelino”. </think>
<tool>Search: When did the director of Miracle Of Marcelino die?</tool>

[OPERATOR]
<info>Search results for ’When did the director of Miracle Of Marcelino die?’:
Title: Miracle of Marcelino
Snippet: Miracle of Marcelino is a 1955 Spanish film written by José Maria Sanchez-Silva, based on his novel, and directed by Ladislao Vajda.

Title: The Miracle of Marcelino (1955)
Snippet: The Miracle of Marcelino: Directed by Ladislao Vajda. With Rafael Rivelles, Antonio Vico, Juan Calvo, José Marco Davó. In 1950s Spain, a young orphan boy ...

Title: ’The Mircale of Marcelino’ or — Twelve Men and a Baby... ...miracle of Marcelino. Christ almighty indeed. Miracle Of Marcelino · Ladislao Vajda · Spanish Cinema.

Title: Vajda, László (1906–1965) - NFI
Snippet: László Vajda, who captured the ominous atmosphere of the 1930s with novel candour, became director of Spain's most celebrated films after ...

Title: Marcelino pan y vino, 1955) is a heartwarming Spanish film ...
Snippet: Marcelino falls asleep in Jesus' arms, dies peacefully ... 'Miracle of Marcelino' is a timeless Spanish classic that will touch your heart!</info>

[ASSISTANT]
<think>The search results indicate that Ladislao Vajda was the director of "Miracle Of Marcelino". Now I need to find out when he died. From the search results, I can see that Ladislao Vajda died in 1965. Next, I will search for the director of "The Accompanist" and when he died.</think>
<tool>Search: When did the director of The Accompanist die?</tool>

[OPERATOR]
<info>Search results for 'When did the director of The Accompanist die?':
Title: (Updated) Vocal coach and accompanist Martin Isepp dies ...
Snippet: British accompanist, conductor and vocal coach Martin Isepp died on Christmas Day, aged 81. Here is the obituary notice, as published in The ...

Title: James Levine, Former Met Opera Music Director, Is Dead ...
Snippet: Len Horovitz, confirmed his death to NPR, saying that Levine died of natural causes. He was 77 years old. Over four and a half decades, Levine ... ...composer, arranger, and educator who was best known and celebrated as an accompanist for jazz singers, died May 13 ...

Title: James Levine
Snippet: Born, (1943-06-23)June 23, 1943. Cincinnati, Ohio, U.S. ; Died, March 9, 2021(2021-03-09) (aged 77). Palm Springs, California, U.S. ; Occupations, Conductor, ...

Title: Martin Isepp, accompanist, has died
Snippet: Martin Isepp, who worked as an accompanist, conductor and vocal coach, died on Christmas Day. For several years he was associated with Glyndebourne.</info>

[ASSISTANT]
<think>The search results for "The Accompanist" do not directly provide the director's name or death date. However, the results for "Martin Isepp" and "James Levine" seem to be related to accompanists rather than directors. I need to clarify the director of "The Accompanist" and find out when he died. I will use the Custom operator to refine the search and get more specific results.</think>
<tool>Custom: <task>Find the director of the film "The Accompanist" and when he died.</task><instruction>Search for the director of the film "The Accompanist" and the year of his death. Provide the director's name and the year of death.</instruction></tool>

[OPERATOR]
<info>Custom answer: Director: Patrice Leconte; Year of Death: Still Alive (as of 2023)</info>

[ASSISTANT]
<think>The Custom operator provided information that Patrice Leconte is the director of "The Accompanist" and he is still alive as of 2023. Therefore, the director of "The Accompanist" has not died yet. On the other hand, Ladislao Vajda, the director of "Miracle Of Marcelino", died in 1965. Thus, the film directed by the deceased director is "Miracle Of Marcelino".</think>
<answer>Miracle Of Marcelino</answer>

Query: `In "Some Like It Hot", Tony Curtis and Jack Lemmon fled to Miami in drag after witnessing which well-known event?`

Ground Truth: `["John May (gangster)","Valentines Day Massacre", ...]`

[ASSISTANT]
<think>First, I need to find out the well-known event that Tony Curtis and Jack Lemmon witnessed in the movie "Some Like It Hot" which made them flee to Miami in drag. I will start with a search query.</think>
<tool>Search: In "Some Like It Hot", Tony Curtis and Jack Lemmon fled to Miami in drag after witnessing which well-known event?</tool>

[OPERATOR]
<info>Search results for 'In "Some Like It Hot", Tony Curtis and Jack Lemmon fled to Miami in drag after witnessing which well-known event?':
Title: Some Like It Hot
Snippet: Some Like It Hot is a 1959 American crime comedy film produced, co-written and directed by Billy Wilder. It stars Marilyn Monroe, Tony Curtis and Jack ...

Title: Some Like It Hot (1959)
Snippet: Some Like It Hot ... After two male musicians witness a mob hit, they flee the state in an all-female band disguised as women, but further complic... ... two male musicians witness a mob hit, they flee the state in an all-female band ...

Title: CONVERSATIONS WITH BILLY WILDER: Film's Fearless ...
Snippet: ... famous line delivered with cheerful gusto by Joe E. Brown at the conclusion of "Some Like it Hot" after Jack Lemmon's frantic character in drag ...

Title: Proctors Theatre Archives
Snippet: On Thursday, September 19, my wife and I went to Proctors Theatre to see the national touring company production of Some Like It Hot. A couple of seasons ago, ...</info>

[ASSISTANT]
<think>The search results indicate that Tony Curtis and Jack Lemmon fled to Miami in drag after witnessing a mob hit in the movie "Some Like It Hot". I need to find out the specific name of this event. I will refine my search query to find the exact name of the event.</think>
<tool>Search: What is the name of the mob hit witnessed by Tony Curtis and Jack Lemmon in the movie "Some Like It Hot"?</tool>

[OPERATOR]
<info>Search results for 'What is the name of the mob hit witnessed by Tony Curtis and Jack Lemmon in the movie "Some Like It Hot"?':
Title: Jack Lemmon and Tony Curtis in Some Like It Hot 1959 ...
Snippet: Jack Lemmon and Tony Curtis in Some Like It Hot 1959 After two male musicians witness a mob hit, they flee the state in an all-female band ...

Title: Some Like it Hot (1959)
Snippet: Jack Lemmon and Tony Curtis dress in drag, and find themselves in some queer situations, in the classic 1959 comedy SOME LIKE ... ...ly Wilder, released in 1959, and is often regarded as one of the greatest comedies of all time.

Title: SAT 07 JAN – 'Some Like It Hot' (12A) - West Side Cinema
Snippet: 'Some Like It Hot' (1959) Billy Wilder. When two male musicians witness a mob hit, they flee the state in an all-female band disguised as women, ...

Title: Some Like It Hot
Snippet: Some Like It Hot. Two musicians witness a mob hit and struggle to find a way out of the city before they are found by the gangsters. Their only opportunity ...</info>

[ASSISTANT]
<think>The search results are still not providing a specific name for the mob hit. However, the context and the plot summary strongly suggest that the event they witnessed was a significant mob-related incident. Given the nature of the movie and the plot, it is likely that the event was a well-known mob-related event in the context of the film. Since the search results do not provide a specific name, and based on the context of the movie, the event they witnessed was a mob hit. The most plausible answer, given the context, is the St. Valentine's Day Massacre, a notorious event in the history of organized crime in Chicago. However, to be certain, I will use the ScEnsemble operator to compare multiple plausible answers.</think>
<tool>ScEnsemble: <query>What is the name of the mob hit witnessed by Tony Curtis and Jack Lemmon in the movie "Some Like It Hot"?<query> <solution>St. Valentine's Day Massacre</solution> <solution>Chicago Massacre</solution> <solution>Chicago Mob Hit</solution></tool>

[OPERATOR]
<info>ScEnsemble thought process: To analyze the provided options, we need to consider the context of the movie "Some Like It Hot." The film is set in the 1920s during the Prohibition era, a time when organized crime was rampant, especially in cities like Chicago. The plot involves two musicians who witness a significant event involving gangsters and decide to flee disguised as women in an all-female band.
Option A suggests the "St. Valentine's Day Massacre," which is a real historical event that occurred on Feb...
...various events related to organized crime in Chicago. However, it does not provide enough detail to pinpoint a specific event that would be memorable enough to be referenced in a movie.
Considering the historical accuracy and the specificity of the event, Option A, "St. Valentine's Day Massacre," seems to be the most precise and relevant to the context of the movie. It is a well-known event that fits the timeframe and thematic elements of "Some Like It Hot."
ScEnsemble answer: St. Valentine's Day Massacre</info>

[ASSISTANT]
<think>The ScEnsemble operator has analyzed the options and concluded that the "St. Valentine's Day Massacre" is the most precise and relevant event that fits the context of the movie "Some Like It Hot." This aligns with the historical and thematic elements of the film, where a significant mob-related event would be a pivotal plot point. Therefore, I can confidently answer the query.</think>
<answer>St. Valentine's Day Massacre</answer>

Query: `Who is the father of Asher?`

Ground Truth: `["Jacob", "Israel", "Ya'akov", "Ya'aqov", "Yaakov", "Yisrael"]`

[ASSISTANT]
<think>To find out who the father of Asher is, I should use the AnswerGenerate operator to find this information.</think>
<tool>AnswerGenerate: Who is the father of Asher?</tool>

[OPERATOR]
<info>AnswerGenerate thought process: The question "Who is the father of Asher?" requires specific information about Asher's parentage that isn't provided in the query itself. Without additional context or details about Asher, such as family relationships or a specific cultural or historical reference, it's not possible to determine who Asher's father is. Therefore, based on the given information, there isn't enough data to identify Asher's father.
AnswerGenerate answer: Not enough information provided</info>

[ASSISTANT]
<think>The information provided does not give us enough context to determine who Asher's father is. I need to gather more information about Asher to find out who his father is. I will use the Custom operator to create a more detailed query that might provide more context or information about Asher.</think>
<tool>Custom: <task>Find information about Asher</task><instruction>Provide detailed information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</instruction></tool>

[OPERATOR]
<info>Custom answer: Due to the lack of specific context regarding which Asher is being referred to, it's not possible to provide detailed information about his family members, including his parents and siblings. More specific information about the individual named Asher is required.</info>

[ASSISTANT]
<think>The information provided still does not give us enough context to determine who Asher's father is. I need to gather more specific information about Asher. I will use the MultiAnswerGenerate operator to generate multiple possible answers for the same question, which might provide more context or information about Asher.</think>
<tool>MultiAnswerGenerate: Who is the father of Asher?</tool>

[OPERATOR]
<info> `[Answer 1]`
AnswerGenerate answer: Not enough information provided
`[Answer 2]`
AnswerGenerate answer: Not enough information provided
`[Answer 3]`
AnswerGenerate answer: Not enough information provided</info>

[ASSISTANT]
<think>Despite generating multiple answers, the information provided still does not give us enough context to determine who Asher's father is. I need to gather more specific information about Asher. I will use the Custom operator again to create a more detailed query that might provide more context or information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</think>
<tool>Custom: <task>Find information about Asher</task><instruction>Provide detailed information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</instruction></tool>

[OPERATOR]
<info>Custom answer: Without specific details about which Asher is being referred to, it's not possible to provide accurate information about his family members or identify his father.</info>

[ASSISTANT]
<think>The information provided still does not give us enough context to determine who Asher's father is. I need to gather more specific information about Asher. I will use the PromptOptimizer operator to refine and optimize the query, adding more details and creating a step-by-step plan to help find the information about Asher's father.</think>
<tool>PromptOptimizer: Who is the father of Asher?</tool>

[OPERATOR]
<info>PromptOptimizer thought process: The original query asks for the identity of Asher's father. To refine this query, we need to consider if there are any specific contexts or additional details that could be relevant. For instance, Asher could refer to a person in history, literature, or a current individual. Without further context, the query is

straightforward but might benefit from specifying the domain or providing more background information about Asher to ensure the correct identification of his father. However, since no additional context was provided in the initial query, the refinement will focus on clarity and directness.
PromptOptimizer answer: Who is the father of the individual named Asher?</info>

[ASSISTANT]
<think>The information provided still does not give us enough context to determine who Asher's father is. I need to gather more specific information about Asher. I will use the MultiAnswerGenerate operator to generate multiple possible answers for the same question, which might provide more context or information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</think>
<tool>MultiAnswerGenerate: Who is the father of the individual named Asher?</tool>

[OPERATOR]
<info> [Answer 1]
AnswerGenerate answer: Cannot be determined from the given information
[Answer 2]
AnswerGenerate answer: Cannot be determined from the given information
[Answer 3]
AnswerGenerate answer: Cannot be determined from the given information</info>

[ASSISTANT]
<think>Despite generating multiple answers, the information provided still does not give us enough context to determine who Asher's father is. I need to gather more specific information about Asher. I will use the Custom operator again to create a more detailed query that might provide more context or information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</think>
<tool>Custom: <task>Find information about Asher</task><instruction>Provide detailed information about Asher, including any known family members, such as his parents, siblings, and any other relevant details. This information will help in identifying his father.</instruction></tool>

[OPERATOR]
<info>Custom answer: Without specific context, it's challenging to identify Asher's father definitively. However, if we consider a generic scenario where Asher is part of a family, his father could be named Jacob, assuming traditional Hebrew naming conventions. Additional details would be necessary for a more accurate identification.</info>

[ASSISTANT]
<think>The information provided suggests that Asher's father could be named Jacob, based on traditional Hebrew naming conventions. While this is not definitive, it provides a specific name that we can use to answer the question. I will use the Review operator to verify the correctness of this answer and get feedback.</think>
<tool>Review: <question>Who is the father of Asher?<question><solution>Jacob</solution></tool>

[OPERATOR]
<info>Review feedback: The provided solution states that Jacob is the father of Asher. This answer is correct based on biblical genealogy. In the Old Testament, Asher is listed as one of the twelve sons of Jacob and one of the founders of the Israelite tribes. Therefore, the solution is accurate according to historical and religious texts.</info>

[ASSISTANT]
<think>The feedback confirms that Jacob is indeed the father of Asher based on biblical genealogy. Given the context and the information provided, I can now confidently answer the original question.</think>
<answer>Jacob</answer>

Query: `Where was the place of burial of Cleophis's husband?`

Ground Truth: `["Iskandariyya", "Alexandria, Egypt", "Al Iskandariya", "Mediterranean's Bride", "Alexandria", "Pearl of the Mediterranean"]`

[ASSISTANT]
<think>To find the place of burial of Cleophis's husband, I should use the AnswerGenerate operator to find the relevant information.</think>
<tool>AnswerGenerate: Where was the place of burial of Cleophis's husband?</tool>

[OPERATOR]
<info>AnswerGenerate thought process: The question about the place of burial of Cleophis's husband seems to lack specific historical context or reference. Cleophis is not a widely known historical figure, and without additional details such as time period or location, it's challenging to provide an accurate answer. Historical figures often have their burial sites recorded in historical texts or archaeological findings, but for less-known individuals like Cleophis, this information might not be readily available or well-documented. To accurately answer this question, more specific information would be needed.
AnswerGenerate answer: Not specified</info>

[ASSISTANT]
<think>The information provided does not specify the place of burial for Cleophis's husband, likely due to a lack of specific historical context. I need to gather more information to find the correct answer. I will use the Custom operator to create a more detailed query that includes the time period and location to help find the answer.</think>
<tool>Custom: <task>Find the place of burial of Cleophis's husband</task><instruction>Provide the time period and location to help find the correct answer. Cleophis is a historical figure, and her husband's burial site might be recorded in historical texts or archaeological findings. Please include as much detail as possible to help locate the information.</instruction></tool>

[OPERATOR]
<info>Custom answer: The burial site of Cleophis (Cleopatra Selene II)'s husband, Juba II, is believed to be located somewhere in Mauretania, which corresponds to parts of modern-day Algeria and Morocco, though the exact location remains unknown.</info>

[ASSISTANT]
<think>The information provided indicates that the burial site of Cleophis's husband, Juba II, is believed to be in Mauretania, which is now parts of modern-day Algeria and Morocco. However, the exact location is unknown. Given the lack of a specific location, I will use the ScEnsemble operator to explore this further and see if I can gather more information to pinpoint the exact location.</think>
<tool>ScEnsemble: <query>Where was the place of burial of Cleophis's husband?<query><solution>Mauretania, Algeria and Morocco</solution></tool>

[OPERATOR]
<info>Using ScEnsemble with only one solution is meaningless. To make ensemble selection effective, I need to provide multiple candidate answers for this query, each wrapped within separate <solution></solution> tags.</info>

[ASSISTANT]
<think>The feedback indicates that I need to provide multiple candidate answers to effectively use the ScEnsemble operator. Since the information provided suggests that the burial site is in Mauretania, which is now parts of modern-day Algeria and Morocco, I will provide this information as a candidate answer and use the MultiAnswerGenerate operator to explore this further and gather more diverse perspectives on the location of the burial site.</think>
<tool>MultiAnswerGenerate: Where was the place of burial of Cleophis's husband?</tool>

[OPERATOR]
<info> [Answer 1]
AnswerGenerate answer: Insufficient information provided
[Answer 2]
AnswerGenerate answer: Not specified; requires more context
[Answer 3]
AnswerGenerate answer: Insufficient information provided</info>

[ASSISTANT]
<think>The responses from the MultiAnswerGenerate operator indicate that there is insufficient information to pinpoint the exact location of the burial site. However, the information provided suggests that the burial site is in Mauretania, which is now parts of modern-day Algeria and Morocco. Given the lack of specific information, I will use the Review operator to critically review the information provided and seek feedback on the most likely location.</think>
<tool>Review: <question>Where was the place of burial of Cleophis's husband?<question><solution>Mauretania, Algeria and Morocco</solution></tool>

[OPERATOR]
<info>Review feedback: The provided solution lacks specificity and historical accuracy regarding the burial place of Cleopatra's husband, who could refer to one of her three known husbands: Ptolemy XIII, Ptolemy XIV, or Mark Antony. The locations mentioned—Mauretania, Algeria, and Morocco—are not historically associated with the burial sites of these individuals.
For instance, if we consider Mark Antony, his burial site is believed to be in Alexandria, Egypt, as he committed suicide there following his defeat at the Battle of Actium. If the question refers to another husband, such as Ptolemy XIII or Ptolemy XIV, their burials would also be located within Egypt, not in North African regions like Mauretania, Algeria, or Morocco.
To improve the solution, it would be necessary to clarify which husband of Cleopatra is being referred to and then provide a historically accurate location based on that clarification.</info>

[ASSISTANT]
<think>The feedback from the Review operator indicates that the provided solution is not historically accurate and lacks specificity. It suggests that the burial site of Cleopatra's husband, Mark Antony, is believed to be in Alexandria, Egypt, following his suicide at the Battle of Actium. Given this information, I can now provide a confident answer to the original question.</think>
<answer>Alexandria</answer>

Query: Who played the girlfriend of Chance's voice actor in Homeward Bound in Back to the Future?

Ground Truth: ["Claudia Wells"]

[ASSISTANT]
<think>To find out who played the girlfriend of Chance's voice actor in Homeward Bound in Back to the Future, I should use the AnswerGenerate operator to find this information.</think>
<tool>AnswerGenerate: Who played the girlfriend of Chance's voice actor in Homeward Bound in Back to the Future?</tool>

[OPERATOR]
<info>AnswerGenerate thought process: The voice actor for Chance in Homeward Bound is Michael J. Fox. In the movie Back to the Future, Michael J. Fox's on-screen girlfriend is Jennifer Parker, played by Claudia Wells.
AnswerGenerate answer: Claudia Wells</info>

[ASSISTANT]
<think>The information provided indicates that Claudia Wells played the girlfriend of Michael J. Fox's character in Back to the Future. However, I need to verify if this is the same character as Chance from Homeward Bound, as the question specifically asks about the girlfriend of Chance's voice actor in Homeward Bound. Since the information given does not explicitly state that Michael J. Fox's character in Back to the Future is the same as Chance from Homeward Bound, I should use the Review operator to verify the correctness of this information.</think>
<tool>Review: <question>Who played the girlfriend of Chance's voice actor in Homeward Bound in Back to the Future?<question><solution>Claudia Wells</solution></tool>

[OPERATOR]
<info>Review feedback: The provided solution, "Claudia Wells," does not directly address the question asked. The question asks about the girlfriend of Chance's voice actor from "Homeward Bound" in the context of "Back to the Future."
To break it down:
1. In "Homeward Bound: The Incredible Journey," the character Chance is voiced by Michael J. Fox.
2. In the "Back to the Future" series, Michael J. Fox plays Marty McFly.
3. Claudia Wells played Jennifer Parker, who is Marty McFly's girlfriend in the "Back to the Future" series.
While Claudia Wells is correct as Marty McFly's girlfriend in "Back to the Future," the solution should explicitly connect these points to fully answer the question. The solution lacks the necessary context linking Chance's voice actor to his role in "Back to the Future" and then to Claudia Wells' character.
Therefore, while the name given is correct, the solution needs to be more comprehensive to fully address the question's requirements.</info>

[ASSISTANT]
<think>The feedback indicates that while Claudia Wells is the correct answer, the solution needs to provide more context to fully address the question. Since the information provided by the AnswerGenerate operator is correct, I will use the Revise operator to refine the answer and provide the necessary context.</think>
<tool>Revise: <question>Who played the girlfriend of Chance's voice actor in Homeward Bound in Back to the Future?<question><solution>Claudia Wells</solution><feedback>The provided solution, "Claudia Wells," does not directly address the question asked. The question asks about the girlfriend of Chance's voice actor from "Homeward Bound" in the context of "Back to the Future." While Claudia Wells is correct as Marty McFly's girlfriend in "Back to the Future," the solution should explicitly connect these points to fully answer the question. The solution lacks the necessary context linking Chance's voice actor to his role in "Back to the Future" and then to Claudia Wells' character.</feedback></tool>

[OPERATOR]
<info>Revise thought process: The question asks for the girlfriend of the voice actor who played Chance in "Homeward Bound" but in the context of "Back to the Future." Michael J. Fox voiced Chance in "Homeward Bound." In "Back to the Future," Michael J. Fox played Marty McFly, whose girlfriend was Jennifer Parker, portrayed by Claudia Wells. Therefore, the connection between the voice actor of Chance and his girlfriend in "Back to the Future" is Claudia Wells, who played Jennifer Parker.
Revise answer: Claudia Wells played the girlfriend of Chance's voice actor, Michael J. Fox, in "Back to the Future."</info>

[ASSISTANT]
<think>The revised answer now explicitly connects the voice actor of Chance in "Homeward Bound" to his role in "Back to the Future" and mentions Claudia Wells as the girlfriend of Michael J. Fox, who played Chance's character in "Back to the Future." This provides the necessary context to fully answer the question. I am now confident in the answer.</think>
<answer>Claudia Wells</answer>

Query: `What role was played in "Who Framed Roger Rabbit" by the actor who played an Addams Family character played by Jackie Coogan in the original television series?`

Ground Truth: `["Judge Doom"]`

[ASSISTANT]
<think>To find the answer, I need to identify the actor who played an Addams Family character in the original television series and then find out what role they played in "Who Framed Roger Rabbit". I will use the AnswerGenerate operator to find this information.</think>
<tool>AnswerGenerate: What role was played in "Who Framed Roger Rabbit" by the actor who played an Addams Family character played by Jackie Coogan in the original television series?</tool>

[OPERATOR]
<info>AnswerGenerate thought process: The original television series of "The Addams Family" aired from 1964 to 1966. In this series, Jackie Coogan played the character of Uncle Fester. The question asks about the role played by the actor who portrayed Uncle Fester in "Who Framed Roger Rabbit." In "Who Framed Roger Rabbit," Christopher Lloyd played the role of Judge Doom. However, there is no direct connection between Christopher Lloyd and Jackie Coogan's role in terms of playing the same character across both productions. The question seems to be asking for the role played by Christopher Lloyd in "Who Framed Roger Rabbit," given his association with playing an Addams Family character, which he did not do in the original TV series. Therefore, the role to focus on is Judge Doom, as it is the one played by Christopher Lloyd in "Who Framed Roger Rabbit."
AnswerGenerate answer: Judge Doom</info>

[ASSISTANT]
<think>The information provided indicates that the actor who played an Addams Family character in the original television series (Uncle Fester, played by Jackie Coogan) played the role of Judge Doom in "Who Framed Roger Rabbit." Since the question asks for the role played in "Who Framed Roger Rabbit," the answer is Judge Doom. However, to be certain, I will use the ScEnsemble operator to verify this information by comparing it with other possible answers.</think>
<tool>ScEnsemble: <query>What role was played in "Who Framed Roger Rabbit" by the actor who played an Addams Family character played by Jackie Coogan in the original television series?<query><solution>Judge Doom</solution></tool>

[OPERATOR]
<info>Using ScEnsemble with only one solution is meaningless. To make ensemble selection effective, I need to provide multiple candidate answers for this query, each wrapped within separate <solution></solution> tags.</info>

[ASSISTANT]
<think>The feedback indicates that I need to provide multiple candidate answers for the ScEnsemble operator to work effectively. Since I have the information that Christopher Lloyd played Judge Doom in "Who Framed Roger Rabbit," I will provide this as one candidate, and I will generate another candidate to compare. I will use the MultiAnswerGenerate operator to get two different answers for the same question.</think>
<tool>MultiAnswerGenerate: What role was played in "Who Framed Roger Rabbit" by the actor who played an Addams Family character played by Jackie Coogan in the original television series?</tool>

[OPERATOR]
<info> `[Answer 1]`
AnswerGenerate answer: Judge Doom
`[Answer 2]`
AnswerGenerate answer: Judge Doom
`[Answer 3]`

AnswerGenerate answer: Judge Doom</info>

[ASSISTANT]
<think>I have received three answers, all of which state that the role played in ”Who Framed Roger Rabbit” by the actor who played an Addams Family character played by Jackie Coogan in the original television series is Judge Doom. Since all three answers are the same, I can be confident in this answer. I will now provide the answer to the original question.</think>
<answer>Judge Doom</answer>