



## WebSailor: Navigating Super-human Reasoning for Web Agent

Kuan Li\*, Zhongwang Zhang\*, Hufeng Yin\*(✉), Liwen Zhang\*, Litu Ou\*, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang(✉), Ming Yan, Pengjun Xie, Fei Huang, Jingren Zhou

Tongyi Lab , Alibaba Group

<https://github.com/Alibaba-NLP/WebAgent>

### Abstract

Transcending human cognitive limitations represents a critical frontier in LLM training. Proprietary agentic systems like DeepResearch have demonstrated superhuman capabilities on extremely complex information-seeking benchmarks such as BrowseComp, a feat previously unattainable. We posit that their success hinges on a sophisticated reasoning pattern absent in open-source models: the ability to systematically reduce extreme uncertainty when navigating vast information landscapes. Based on this insight, we introduce WebSailor, a complete post-training methodology designed to instill this crucial capability. Our approach involves generating novel, high-uncertainty tasks through structured sampling and information obfuscation, RFT cold start, and an efficient agentic RL training algorithm, Duplicating Sampling Policy Optimization (DUPO). With this integrated pipeline, WebSailor significantly outperforms all open-source agents in complex information-seeking tasks, matching proprietary agents' performance and closing the capability gap.

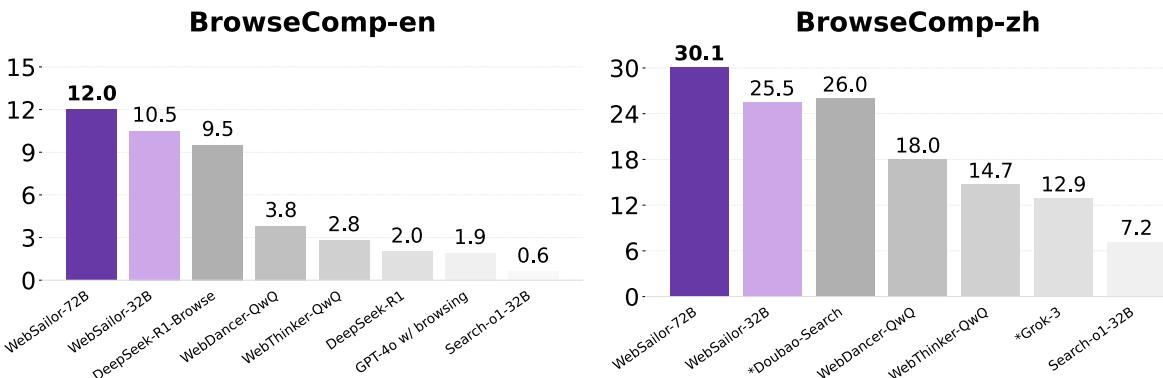


Figure 1: Performance on the BrowseComp-en/zh benchmarks. DeepSeek-R1-Browse is DeepSeek-R1 equipped with browsing tools via the ReAct framework, sharing the same implementation as our model, WebSailor. Doubao-Search and Grok-3 are proprietary web-based products (marked by \*). The result for GPT-4o with browsing is taken from OpenAI's official publication.

\*Equal Core Contributors. Kuan Li, Zhongwang Zhang, and Hufeng Yin are project leaders.

✉Corresponding author. {yinhuifeng.yhf, yongjiang.yj}@alibaba-inc.com



## WebSailor: Web Ajanı için Süper İnsan Akıl Yürütmeye Gezinmek

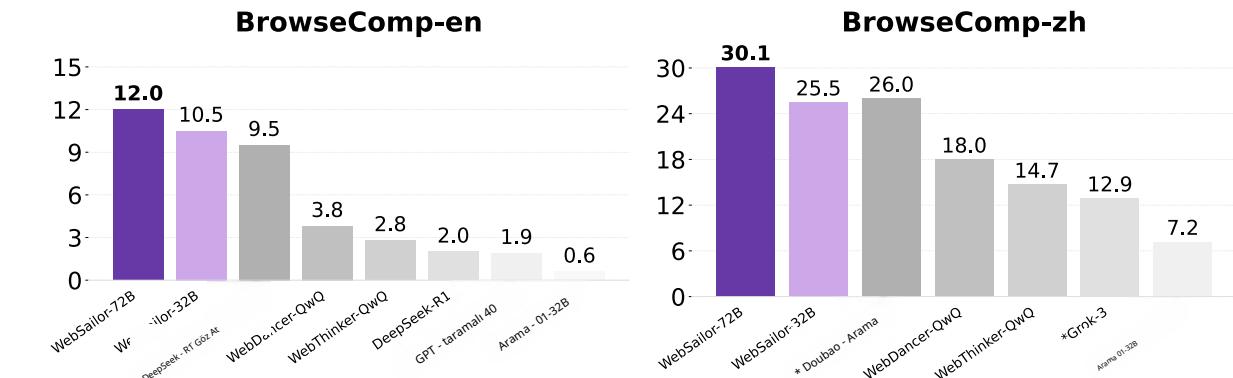
Kuan Li\*, Zhongwang Zhang\*, Hufeng Yin\*(✉), Liwen Zhang\*, Litu Ou\*, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang(✉), Ming Yan, Pengjun Xie, Fei Huang, Jingren Zhou

Tongyi Lab , Alibaba Grubu

<https://github.com/Alibaba-NLP/WebAgent>

### Özet

İnsan bilişsel sınırlamalarını aşmak, Büyük Dil Modeli (LLM) eğitimi için kritik bir sınırdır. DeepResearch gibi tescilli ajantik sistemler, daha önce ulaşılamayan BrowseComp gibi son derece karmaşık bilgi arama kıyaslamalarında süper insan yetenekleri sergilemiştir. Başarılarının, açık kaynak modellerde bulunmayan sofistik bir akıl yürütme örüntüsüne dayandığını savunuyoruz: geniş bilgi ortamlarında gezinirken aşırı belirsizliği sistematik olarak azaltma yeteneği. Bu içgüdüye dayanarak, bu kritik yeteneği kazandırmak için tasarlanmış eksiksiz bir eğitim sonrası metodoloji olan WebSailor'ı tanıttık. Yaklaşımımız, yapılandırılmış örneklemeye ve bilgi gizleme yoluyla yeni, yüksek belirsizlik içeren görevler üretmeyi, RFT soğuk başlatmayı ve verimli bir ajansal RL eğitim algoritması olan Duplicating Sampling Policy Optimization (DUPO)'yu içermektedir. Bu entegre boru hattı ile WebSailor, karmaşık bilgi arama görevlerinde tüm açık kaynaklı ajanları önemli ölçüde geride bırakarak, tescilli ajanların performansına ulaşmakta ve yetenek açığını kapatmaktadır.



Şekil 1: BrowseComp-en/zh kıyaslama testlerindeki performans. DeepSeek-R1-Browse, ReAct çerçevesi aracılığıyla tarama araçlarıyla donatılmış DeepSeek-R1'dir ve modelimiz WebSailor ile aynı uygulamayı paylaşmaktadır. Doubao-Search ve Grok-3, tescilli web tabanlı ürünlerdir (\* ile işaretlenmiştir). Tarama özellikle GPT-4o'nun sonucu OpenAI'nın resmi yayınından alınmıştır.

\*Eşit Çekirdek Katkıda Bulunanlar. Kuan Li, Zhongwang Zhang ve Hufeng Yin proje liderleridir.

Sorumlu yazar. {yinhuifeng.yhf, yongjiang.yj}@alibaba-inc.com

## 1 Introduction

Information seeking, the fundamental human drive to resolve uncertainty, has been revolutionized by the internet (Wilson, 1999; Jurado et al., 2015). Yet, human ability to navigate this vast digital landscape is constrained by cognitive limits: finite memory, fragile attention, and an inability to pursue multiple exploratory paths in parallel. Leading proprietary agentic systems, such as Deep Research (OpenAI, 2025a), show that Large Language Model (LLM) agents can transcend these human limitations. Their superhuman performance on complex web benchmarks like BrowseComp-en/zh (Wei et al., 2025; Zhou et al., 2025) stems from sophisticated reasoning—internal or tool-mediated—that systematically reduces uncertainty (Kapoor et al., 2024; Huang et al., 2023).

However, instilling these advanced reasoning capabilities in open-source agents remains an unsolved problem. As shown in Fig. 1, existing open-source LLMs and web agents exhibit near-zero accuracy on BrowseComp-en (Wu et al., 2025a; Li et al., 2025c;b; Song et al., 2025). This stark performance gap arises because current training paradigms focus on what we classify as Level 1 and 2 tasks: problems with either low uncertainty (e.g., single-search) or a clear, structured path to resolution (e.g., standard multi-hop QA). These datasets do not expose models to the Level 3 challenges that dominate complex benchmarks—scenarios demanding robust compositional generalization (Wiedemer et al., 2023) over intricate information landscapes with no predefined solution path. Consequently, models fail to develop the complex, multi-step reasoning required to navigate them.

To elicit these superhuman reasoning patterns, we generate training data characterized by high and hard-to-reduce intrinsic uncertainty. Our primary mechanism involves sampling subgraphs from interconnected knowledge structures generated by random walks across real-world websites. From a compositional generalization perspective (Google, 2020), these subgraphs present novel combinations of known entities and relationships, forcing the model to reason about previously unseen compositions and pushing it beyond simple heuristics. This process generates a diverse array of intricate, emergent structures that are difficult to pre-define, compelling the model to develop reasoning processes that may transcend established human patterns.

We further amplify task difficulty using carefully designed information obfuscation techniques, which directly increase initial ambiguity. The combination of structural complexity and informational ambiguity creates tasks that demand exceptionally sophisticated reasoning. For instance, some of our generated questions are so challenging that even powerful proprietary models like o3 (OpenAI, 2025c) require up to 40 tool calls to arrive at a solution, underscoring the extreme uncertainty reduction involved.

After obtaining QAs, a key challenge is acquiring full supervision. While powerful open-source Large Reasoning Models (LRMs) like QwQ (Qwen Team, 2025) and DeepSeek-R1 (Guo et al., 2025) can solve some complex QAs, their native reasoning outputs are unsuitable for direct fine-tuning. These models exhibit highly stylized and verbose thought processes that, if imitated, could restrict the trainee agent’s ability to develop its own flexible, exploratory strategies. Furthermore, in long-horizon web tasks requiring dozens of tool calls Li et al. (2025a), their lengthy reasoning chains quickly overwhelm the context window, leading to performance degradation and poor readability (Yin et al., 2025). To overcome this, we propose a novel approach: we leverage these open-source LRMs to generate successful action-observation traces, but then reconstruct the reasoning. By inferring concise, action-oriented thoughts for each step, we create a clean, effective supervision signal that captures the solution logic without inheriting stylistic or verbosity-related drawbacks.

In terms of training process optimization, although recent studies suggest skipping SFT (Guo et al., 2025; Chen et al., 2025; Hu et al., 2025), we demonstrate that a modest rejection sampling fine-tuning (RFT) cold start is indispensable for web agents navigating such complex tasks. On one hand, RL rewards for these scenarios are extremely sparse, often yielding near-zero feedback initially. On the other hand, our approach does not heavily rely on distillation; a minimal cold start with just over  $2k$  high-quality

## 1 Giriş

Belirsizliği gidermeye yönelik temel güdüsü olan bilgi arayışı, internet tarafından devrim niteliğinde bir değişime uğramıştır (Wilson, 1999; Jurado et al., 2015). Ancak, insanın bu geniş dijital manzara içinde gezinme yeteneği bilişsel sınırlamalarla kısıtlıdır: sınırlı hafıza, kırılgan dikkat ve birden fazla keşif yolunu paralel olarak takip edememe. Deep Research (OpenAI, 2025a) gibi onde gelen tescilli ajans sistemleri, Büyük Dil Modeli (LLM) ajanlarının bu insan sınırlamalarını aşabileceğini göstermektedir. BrowseComp-en/zh (Wei et al., 2025; Zhou et al., 2025) gibi karmaşık web kıyaslamalarında göstergeleri süper insan performansı, belirsizliği sistematik olarak azaltan gelişmiş akıl yürütmeden —dahili veya araç aracı— kaynaklanmaktadır (Kapoor et al., 2024; Huang et al., 2023).

Ancak, bu gelişmiş akıl yürütme yeteneklerini açık kaynaklı ajanlara kazandırmak çözülmemiş bir sorun olmaya devam etmektedir. Şekil 1'de gösterildiği gibi, mevcut açık kaynaklı BDM'ler ve web ajanları BrowseComp-en üzerinde sıfıra yakın doğruluk sergilemektedir (Wu et al., 2025a; Li et al., 2025c;b; Song et al., 2025). Bu belirgin performans farkı, mevcut eğitim paradigmalarının Seviye 1 ve 2 olarak sınıflandırıldığımız görevlere odaklanmasından kaynaklanmaktadır: yani ya düşük belirsizliğe sahip (örn. tek arama) ya da çözüme giden açık, yapılandırılmış bir yola sahip (örn. standart çok adımlı QA) problemlere. Bu veri kümeleri, modelleri karmaşık kıyaslamalarda baskın olan Seviye 3 zorluklarına maruz bırakmaz; bu zorluklar, önceden tanımlanmış bir çözüm yolu olmayan karmaşık bilgi ortamlarında sağlam bir bileşimsel genelleme (Wiedemer vd., 2023) gerektiren senaryoları kapsar. Sonuç olarak, modeller bu ortamları yönlendirmek için gereken karmaşık, çok adımlı akıl yürütmemektedir.

Bu süper insan akıl yürütme kalıplarını ortaya çıkarmak amacıyla, yüksek ve azaltılması zor içsel belirsizlikle karakterize edilen eğitim verileri üretiyoruz. Temel mekanizmamız, gerçek dünya web sitelerinde rastgele yürüyüşlerle oluşturulan birbirile bağlı bilgi yapılarından alt grafikler örneklemeyi içermektedir. Bileşimsel genelleme perspektifinden (Google, 2020) bakıldığından, bu alt grafikler bilinen varlıkların ve ilişkilerin yeni kombinasyonlarını sunarak, modeli daha önce görülmemiş bileşimler hakkında akıl yürütmeye zorlar ve onu basit sezgisel yöntemlerin ötesine iter. Bu süreç, önceden tanımlanması güç çeşitli karmaşık, öngörelemeyen yapılar üretir ve modeli, yerleşik insan kalıplarının ötesine geçebilecek akıl yürütme süreçleri geliştirmeye zorlar.

Başlangıçtaki belirsizliği doğrudan artırın, dikkatle tasarlanmış bilgi gizleme teknikleri kullanarak görev zorluğunu daha da artırıyoruz. Yapısal karmaşıklık ve bilgi belirsizliğinin birleşimi, olağanüstü derecede sofistike akıl yürütme gerektiren görevler ortaya çıkarır. Örneğin, oluşturduğumuz bazı sorular o kadar zorlayıcıdır ki, o3 (OpenAI, 2025c) gibi güçlü tescilli modeller dahi bir çözüme ulaşmak için 40'a kadar araç çağrıları gerektirmekte, bu da içerdeği aşırı belirsizlik azaltımının altını çizmektedir.

Soru-Cevap (QA) çiftlerini elde ettikten sonra, ana zorluk tam denetim sağlamaktır. QwQ (Qwen Ekibi, 2025) ve DeepSeek-R1 (Guo vd., 2025) gibi güçlü açık kaynak Büyük Akıl Yürütme Modelleri (LRM'ler) bazı karmaşık QA'ları çözebilse de, doğal akıl yürütme çıktıları doğrudan ince ayar için uygun değildir. Bu modeller, taklit edildiği takdirde, eğitimdeki ajanın kendi esnek, keşifçi stratejilerini geliştirme yeteneğini kısıtlayabilecek, oldukça stilize edilmiş ve ayrıntılı düşünme süreçleri sergilemektedir. Ayrıca, onlarca araç çağrıları gerektiren uzun ufuklu web görevlerinde (Li vd., 2025a), bu modellerin uzun muhakeme zincirleri bağlam penceresini hızla aşarak performans düşüşüne ve düşük okunabilirliğe yol açmaktadır (Yin vd., 2025). Bunu aşmak için yeni bir yaklaşım öneriyoruz: Bu açık kaynaklı LRM'leri başarılı eylem-gözlem izleri oluşturmak için kullanıyor, ancak ardından muhakemeyi yeniden yapılandıryoruz. Her adım için kısa, eylem odaklı düşünceler çıkararak, stilistik veya ayrıntı ile ilgili dezavantajları devralmadan çözüm mantığını yakalayan temiz ve etkili bir denetim sinyali oluşturuyoruz.

Eğitim süreci optimizasyonu açısından, son çalışmalar SFT'yi atlamayı önerse de (Guo vd., 2025; Chen vd., 2025; Hu vd., 2025), bu tür karmaşık görevlerde gezinen web ajanları için mütevazı bir ret örneklemesi ince ayar (RFT) soğuk başlangıcının vazgeçilmez olduğunu gösteriyoruz. Bir yandan, bu senaryolar için RL ödülleri son derece seyrek ve başlangıçta genellikle sıfır yakını geri bildirim sağlamaktadır. Öte yandan, yaklaşımız büyük ölçüde damıtma dayanmamaktadır; yalnızca  $2k$  yüksek kaliteli

examples proves effective. The RL training of agents for such tasks is extremely slow due to multi-turn reasoning and heavy tool use. To address this, we propose Duplicating Sampling Policy Optimization (DUPO), which incorporates two dynamic sampling strategies—one before training and one during training—to improve both effectiveness and efficiency.

Our family of WebSailor models (3B, 7B, 32B, and 72B) outperform all open-source models and agentic methods on BrowseComp-en/zh, and also surpass proprietary LRM's such as Grok-3 ([x.ai, 2025](#)) and DouBao ([DouBao, 2025](#)) when they are combined with browsing capabilities, as shown in Fig. 1. Additionally, we find that post-training based on complex, uncertainty-driven reasoning patterns exhibits downward compatibility, achieving promising performance on simpler tasks such as GAIA ([Mialon et al., 2023](#)), XBench-DeepSearch ([Xbench-Team, 2025](#)), and SimpleQA ([OpenAI, 2025d](#)).

## 2 Problem Definition

We adopt the ReAct ([Yao et al., 2023](#)) as the agent's framework. Upon receiving a question, the agent performs several iterations of Thought-Action-Observation. Specifically, in each iteration, based on the existing context, the LLM generates a Thought and executes a parsable Action (tool call), then awaits the environment to return an Observation. In WebTraverseX, the action space consists of generating final answer and two tools, search and visit, which correspond to invoking a search engine with several queries and accessing several webpages via URLs to retrieve their content, respectively. The details of these two tools are provided in the Appendix A.1 The observation returned by the search action consists of 10 titles, snippets, and their corresponding URLs for each search query. In contrast, the observation of the visit action is a summary of the webpages, tailored to the "goal" specified in the LLM's action. The iteration terminates when the LLM selects "final answer" as the action. A complete trajectory with  $T$  iterations can be defined as:

$$\mathcal{H}_T = (\tau_0, a_0, o_0, \dots, \tau_i, a_i, o_i, \dots, \tau_T, a_T), \quad (1)$$

where  $\tau_i, a_i, o_i$  represent thought, action, and observation in the  $i$ -th round, respectively. At step  $t$ , the thought  $\tau_t$  and  $a_t$  are sampled from a policy based on all previous context, i.e.,  $\pi(a, t | \mathcal{H}_{t-1})$ .

Completing multi-hop QA ([Yang et al., 2018; Ho et al., 2020](#)) typically requires only one or two rounds of ReAct, as the actions at each step are quite clear and do not involve much strategic planning. In stark contrast, BrowseComp immerses the agent in a vast, unstructured information space where the solution path is not predefined. A naive, brute-force search is computationally infeasible, potentially requiring thousands of tool calls that would overwhelm the context window of any modern LLM. Success, therefore, hinges not on following a simple script, but on executing a highly adaptive search strategy. The agent must dynamically synthesize partial information, prune unpromising exploratory paths, and integrate disparate facts to converge on a solution. Compressing this combinatorially vast search space into a tractable trajectory of a few dozen steps requires a sophisticated chain of thought ([Wei et al., 2022](#)). It is precisely this process of strategic navigation and synthesis that exemplifies the complex, superhuman reasoning patterns this work seeks to elicit and model.

## 3 Large-scale Training Data Synthesis for Complex Reasoning

In this section, we present our training data construction from two perspectives: QA construction and reasoning trajectory generation.

### 3.1 SailorFog-QA: Scalable Graph-Synthesized QA

The reasoning patterns required to answer a question are dictated by its intrinsic uncertainty and the complexity of reducing that uncertainty. As shown in Fig. 2, we classify information-seeking QAs into

örnekle yapılan minimal bir soğuk başlangıç etkili olduğunu kanıtlamaktadır. Çok turlu akıl yürütme ve yoğun araç kullanımı nedeniyle bu tür görevler için ajanların RL eğitimi son derece yavaştır. Bu durumu ele almak için, hem etkinliği hem de verimliliği artırmak amacıyla biri eğitimden önce diğerini eğitim sırasında olmak üzere iki dinamik örnekleme stratejisini birleştiren Duplicating Sampling Policy Optimization (DUPO) yöntemini öneriyoruz.

WebSailor model ailemiz (3B, 7B, 32B ve 72B), BrowseComp-en/zh üzerinde tüm açık kaynak modelleri ve ajan tabanlı yöntemlerden daha iyi performans göstermektedir; ayrıca Şekil 1'de gösterildiği gibi tarama yetenekleriyle birleştirildiğinde Grok-3 ([x.ai, 2025](#)) ve DouBao ([DouBao, 2025](#)) gibi tescilli LRM'leri de geride bırakmaktadır. Ek olarak, karmaşık, belirsizlik odaklı akıl yürütme kalıplarına dayalı son eğitim sürecinin geriye dönük uyumluluk gösterdiğini ve GAIA ([Mialon et al., 2023](#)), XBench-DeepSearch ([Xbench-Team, 2025](#)) ve SimpleQA ([OpenAI, 2025d](#)) gibi daha basit görevlerde umut vadeden bir performans sergilediğini tespit etti.

## 2. Problem Tanımı

Aracının çerçevesi olarak ReAct'ı ([Yao et al., 2023](#)) benimsiyoruz. Bir soru aldıktan sonra, aracı Düşünce-Eylem-Gözlem döngüsünün birkaç iterasyonunu gerçekleştirir. Özellikle, her iterasyonda, mevcut bağlama dayanarak, LLM bir Düşünce üretir ve ayırtılabilir bir Eylem (araç çağrı) yürütür, ardından ortamın bir Gözlem döndürmesini bekler. WebTraverseX'te, eylem alanı nihai cevabı ve iki aracı, search ve ziyaret'i oluşturmaktan oluşur; bunlar sırasıyla birkaç sorgu ile bir arama motorunu çağırma ve içeriklerini almak için URL'ler aracılığıyla birkaç web sayfasına erişmeye karşılık gelir. Bu iki aracın ayrıntıları Ek A.1'de sunulmuştur. Arama eylemi tarafından döndürülen gözlem, her arama sorgusu için 10 başlık, özet ve bunlara karşılık gelen URL'lerden oluşur. Buna karşılık, ziyaret eyleminin gözlemi, LLM'nin eyleminden belirtilen "hedefe" göre uyarlanmış web sayfalarının bir özetidir. Yineleme, LLM'nin eylem olarak "nihai yanıt"ı seçmesiyle sonlanır. T yinelemeye sahip tam bir yörunge şıkkıde tanımlanabilir:

$$\mathcal{H}_T = (\tau_0, a_0, o_0, \dots, \tau_i, a_i, o_i, \dots, \tau_T, a_T), \quad (1)$$

burada  $\tau_i, a_i, o_i$  sırasıyla  $i$ -inci turdaki düşünceyi, eylemi ve gözlemi temsil eder.  $t$  adımında,  $\tau_t$  düşünceyi ve  $a_t$  eylemi, önceki tüm bağlama dayalı bir politikadan örneklenir, yani  $\pi(a_t | H_{t-1})$ .

Çok adımlı QA'yı ([Yang ve diğerleri, 2018; Ho ve diğerleri, 2020](#)) tamamlamak genellikle yalnızca bir veya iki ReAct turu gereklidir; zira her adımdaki eylemler oldukça belirgindir ve yoğun stratejik planlama içermez. Bunun aksine, BrowseComp ajanı, çözüm yolunun önceden tanımlanmadığı geniş, yapılandırılmış bir bilgi alanına daldırır. Naif, kaba kuvvet araması hesaplama açısından uygulanamaz olup, potansiyel olarak modern bir LLM'nin bağlam penceresini aşacak binlerce araç çağrı gereklidir. Bu nedenle başarı, basit bir komut dosyasını takip etmekten ziyade, son derece uyarlanabilir bir arama stratejisi yürütmeye bağlıdır. Ajan, kısmi bilgiyi dinamik olarak sentezlemeli, umut vaat etmeyen keşif yollarını eleyerek budamalı ve bir çözüme yakınsamak için farklı gerçekleri birleştirmelidir. Bu kombinatoryal olarak geniş arama uzayını, birkaç düzine adım-dan oluşan izlenebilir bir yörungeye sıkıştırmak, sofistike bir düşünce zinciri ([Wei ve diğerleri, 2022](#)) gereklidir. Bu çalışma tam da bu stratejik gezinme ve sentez sürecini, karmaşık, süper insan akıl yürütme kalıplarını ortaya çıkarmayı ve modellemeyi amaçlamaktadır.

## 3 Karmaşık Akıl Yürütme için Büyük Ölçekli Eğitim Verisi Sentezi

Bu bölümde, eğitim verisi oluşturma sürecimizi iki bakış açısından sunuyoruz: QA oluşturma ve akıl yürütme yörungesi üretimi.

### 3.1 SailorFog-QA: Ölçeklenebilir Grafik Sentezli QA

Bir soruya yanıtlamak için gerekli akıl yürütme örüntülerini, sorunun içsel belirsizliği ve bu belirsizliği azaltmanın karmaşıklığı tarafından belirlenir. Şekil 2'de gösterildiği gibi, bilgi arama QA'larını aşağıdaki şekilde sınıflandırırız:

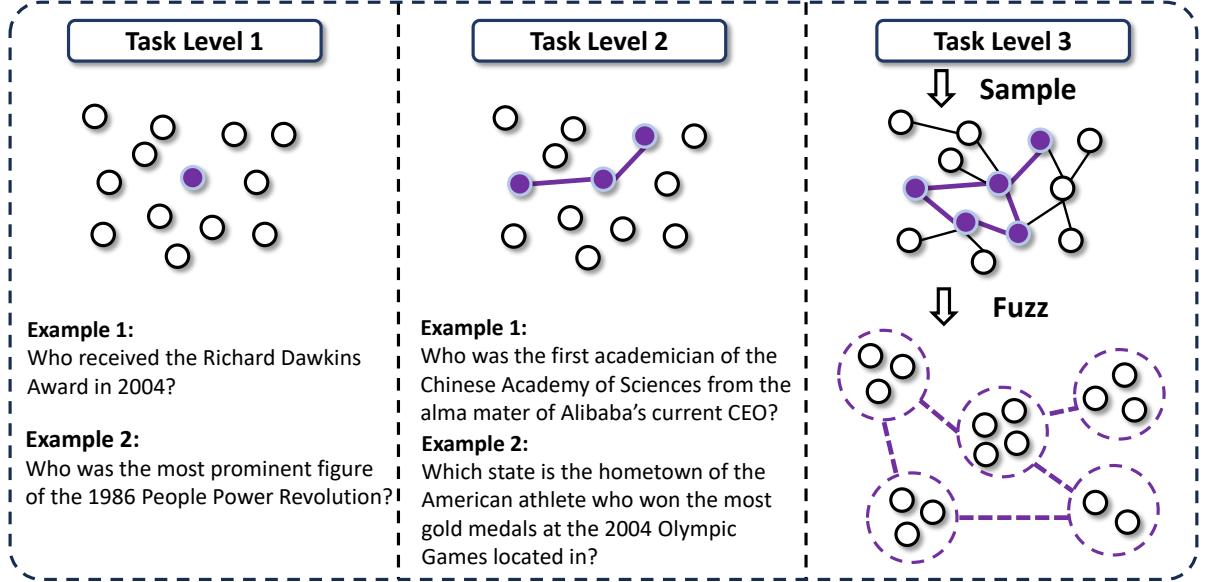


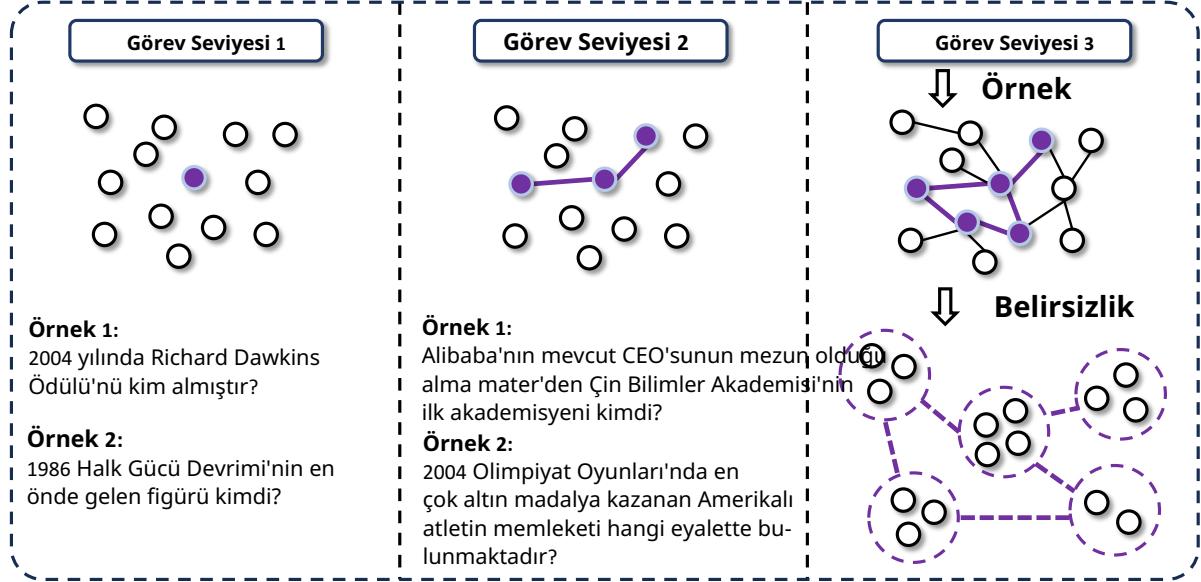
Figure 2: Information seeking tasks can be categorized into three levels. Level 1 features a relatively simple logical structure and can be answered directly or with a single, straightforward tool invocation. Level 2 resembles multi-hop questions, where solutions are obtained through a fixed sequence of reasoning steps. Level 3 exhibits the most complex and variable topology after obfuscation, making it difficult to define manually, and its inherent uncertainty is challenging to reduce.

three levels based on these two dimensions.

- **Level 1: tasks exhibit low uncertainty that is easily reduced.** These include questions answerable by the model’s internal knowledge or through a single, straightforward web search.
- **Level 2: tasks, such as multi-hop QA, present high initial uncertainty but follow a clear path to resolution.** Even with many steps, the entities are linked by well-defined logic, allowing uncertainty to be systematically reduced through a structured sequence of actions.
- **Level 3: the focus of our work, involves problems with both high uncertainty and high difficulty in its reduction.** Here, entities are coupled in complex, emergent ways, lacking a pre-defined reasoning path. Solving these problems demands creative exploration and novel reasoning patterns that are difficult to specify manually.

**Constructing the Structural Foundation for Hard-to-Reduce Uncertainty** To generate Level 3 tasks, we first construct a complex informational landscape where uncertainty is inherently difficult to reduce. Our process, inspired by random walks, builds knowledge graphs with emergent, non-linear structures. We begin by seeding the graph with a fuzzy entity retrieved from Wikidata’s SPARQL service to ensure a challenging starting point. Using simulated web browsing, we gather unstructured text and features about this entity from the internet. From this raw information, we extract related entities and the relationships connecting them, forming the initial nodes and edges. The crucial step is the iterative expansion: we probabilistically select existing nodes and seek out new, distinct entities to connect. This stochastic process discourages simple linear chains (characteristic of Level 2 tasks) and instead fosters a densely interconnected graph with intricate, overlapping relational paths. The resulting graph serves as a structural foundation for problems that lack a pre-defined reasoning path, compelling an agent to navigate a complex web of information rather than follow a straight line.

**Generating High-Uncertainty Questions via Subgraph Sampling and Obfuscation** With these complex graphs as a foundation, we generate questions characterized by high initial uncertainty. This is



Şekil 2: Bilgi arama görevleri üç seviyeye ayrılabilir. Seviye 1, nispeten basit bir mantıksal yapıya sahiptir ve doğrudan veya tek, anlaşılır bir araç çağrı ile yanıtlanabilir. Seviye 2, çözümlerin belirli bir akıl yürütme adımları dizisi aracılığıyla elde edildiği çok adımlı sorulara benzer. Seviye 3, gizlendikten sonra en karmaşık ve değişken topolojiyi sergiler; bu durum, manuel olarak tanımlamasını zorlaştırmakta ve doğal belirsizliğinin azaltılması güç olmaktadır.

bu iki boyuta dayalı üç seviye.

- **Seviye 1: kolayca azaltılabilen düşük belirsizlik gösteren görevler.** Bunlar, modelin dahili bilgisi veya tek, basit bir web araması yoluyla yanıtlanabilecek soruları içerir.
- **Seviye 2: çok adımlı Soru-Cevap (QA) gibi görevler, başlangıçta yüksek belirsizlik sunar ancak çözüme giden net bir yol izler.** Birçok adımda bile, varlıklar iyi tanımlanmış mantıkla birbirine bağlıdır; bu da belirsizliğin yapılandırılmış bir eylemler dizisi aracılığıyla sistematik olarak azaltılmasına olanak tanır.
- **Seviye 3: çalışmamızın odak noktası olan, hem yüksek belirsizliğe hem de bunun azaltılmasında yüksek zorluğa sahip sorunları içerir.** Burada, varlıklar karmaşık, ortaya çıkan yollarla birleşiktir ve önceden tanımlanmış bir akıl yürütme yolu bulunmamaktadır. Bu sorunların çözümü, manuel olarak tanımlanması güç yaratıcı keşif ve yenilikçi akıl yürütme örüntülerini gerektirmektedir.

**Azaltılması Zor Belirsizlik İçin Yapısal Temelin Oluşturulması** Seviye 3 görevlerini oluşturmak amacıyla, öncelikle belirsizliğin doğası gereği indirgenmesinin zor olduğu karmaşık bir bilgi ortamı inşa ederiz. Rastgele yürüyüşlerden esinlenen sürecimiz, ortaya çıkan, doğrusal yapılarla bilgi grafikleri oluşturur. Zorlu bir başlangıç noktası sağlamak üzere, grafiği Wikidata’nın SPARQL hizmetinden elde edilen belirsiz bir varlık ile tohumlayarak işe başlarız. Simülle edilmiş web taramasını kullanarak, bu varlık hakkındaki yapılandırılmış metinleri ve özellikleri internetten toplarız. Bu ham bilgiden, ilgili varlıklar ve bunları birbirine bağlayan ilişkileri çıkararak başlangıç düğümlerini ve kenarlarını oluştururuz. Kritik adım, yinelemeli genişlemedir: Mevcut düğümleri olasılıksal olarak seçer ve bağlanacak yeni, farklı varlıklar ararız. Bu stokistik süreç, basit doğrusal zincirleri (Seviye 2 görevlerinin karakteristik özelliği) teşvik etmek yerine, karmaşık, örtüsen ilişkisel yollara sahip, yoğun bir şekilde birbirine bağlı bir grafik yapısı oluşturur. Ortaya çıkan grafik, önceden tanımlanmış bir muhakeme yolu olmayan problemler için yapısal bir temel görevi görerek, bir ajanı düz bir çizги takip etmek yerine karmaşık bir bilgi ağında gezmeye mecbur bırakır.

**Alt Grafik Örnekleme ve Gizleme Yoluyla Yüksek Belirsizlikli Sorular Üretme** Bu karmaşık grafikleri temel alarak, yüksek başlangıç belirsizliğiyle karakterize edilen sorular üretiyoruz. Bu,

achieved by sampling subgraphs with diverse topologies, each representing a unique constellation of coupled entities and relations. We then formulate a question and an answer based on the subgraph. Critically, we introduce ambiguity through deliberate information obfuscation. Instead of presenting clear facts, we obfuscate features and relationships within the question. For example, a precise date is transformed into a vague period ("in the early 2010s"), a name is partially masked ("an institution founded by someone with the initial 'F'"), or a quantitative attribute is described qualitatively ("a market share of less than 1%"). This obfuscation directly increases the initial uncertainty, forcing the agent to reason, compare, and synthesize information rather than simply executing a lookup. We name our synthetic training data **SailorFog-QA**, which has three key advantages:

- The data is grounded in the real-world internet, mirroring the challenges agents face in practice.
- The diverse subgraph topologies naturally produce problems requiring a spectrum of complex reasoning patterns, from multi-step deduction to compositional and comparative analysis.
- The approach is highly scalable, as the number of potential subgraphs—and thus challenging questions—grows non-linearly with the graph size, enabling efficient large-scale data synthesis.

To illustrate the characteristics of our generated Level 3 tasks, two examples are presented below. These questions epitomize our methodology: they feature multiple, intricately coupled entities and deliberately obfuscated information, such as vague time references ("around the mid-5th century", "early 21st century") and non-specific descriptors ("a prominent South American capital", "a respected arts institution"). This combination of structural complexity and informational ambiguity creates a high degree of initial uncertainty that is exceptionally difficult to reduce. In fact, our manual evaluations confirm that these problems are intractable for human researchers under typical time constraints (e.g., within two hours), as they lack clear starting points for search and require extensive, non-linear exploration. Further details on our QA generation process are available in Appendix A.2.

#### Examples of Generated Questions

**Question:** *There was an early Christian poetic hymn composed by a late antique writer who passed away around the mid-5th century. The year of this writer's death coincides with the last year of a scientific chronology that reconstructs environmental conditions from several centuries before the modern era. What is the name of this chronology?*

**Answer:** *Estimated Tree-Ring Chronology: 300-450 A.D.*

**Question:** *A musical piece closely associated with a prominent South American capital features lyrics written by a notable figure who was later recognized with a distinguished local civic honor in the early 21st century. The composition's melody was created by a musician who received formal training at a respected arts institution in western Colombia. What is the name of this musical piece?*

**Answer:** *the Rue de Rivoli*

### 3.2 Reconstructing Reasoning from Expert LRM Trajectories

Having synthesized complex QA pairs, the next challenge is to generate corresponding solution trajectories for cold-start supervision. While powerful open-source LRMs like QwQ-32B (Qwen Team, 2025) can provide some correct trajectories, directly using their full output for fine-tuning is counterproductive. We identify two critical issues:

- **Stylistic Contamination:** These LRMs possess strong, often verbose, stylistic priors in their reasoning. Directly fine-tuning on these outputs can be overly prescriptive, stifling the agent's ability to develop its own exploratory strategies and generalize to unseen problems.
- **Context Overload:** The verbosity of their reasoning chains is a practical barrier for complex web agent tasks. A trajectory with dozens of tool calls can easily generate a history that exceeds the context limits, degrading performance and making the reasoning process intractable.

çeşitli topolojilere sahip alt grafikler örneklenerek başarılar; her biri, birbirine bağlı varlıkların ve ilişkilerin benzersiz bir takımı自己を表すことを示す。Daha sonra alt grafiğe dayanarak bir soru ve bir yanıt formüle ederiz。Önemli olarak, kasıtlı bilgi gizleme yoluyla belirsizlik oluşturuyoruz。Net gerçekler sunmak yerine, sorudaki özellikleri ve ilişkileri gizliyoruz。Örneğin, kesin bir tarih belirsiz bir dönemde ("2010'ların başlarında") dönüştürüldür, bir isim kısmen maskelenir ("F' baş harfli biri tarafından kurulan bir kuruş") veya nice bir özellik niteliksel olarak tanımlanır ("%1'den az pazar payı")。Bu belirsizleştirme, başlangıçtaki belirsizliği doğrudan artırarak ajanı sadece basit bir arama yapmak yerine bilgi üzerinde akıl yürütmeye, karşılaşmaya ve sentezlemeye zorlar。Sentetik eğitim verilerimizi **SailorFog-QA** adlandırdık ve bunun üç temel avantajı bulunmaktadır:

- Veriler, ajanların pratikte karşılaştığı zorlukları yansitan gerçek dünya internetine dayalıdır。
- Çeşitli alt grafik topolojileri, çok adımlı çıkarımdan kompozisyonel ve karşılaştırmalı analize kadar uzanan karmaşık akıl yürütme kalıpları gerektiren soruları doğal olarak üretir。
- Yaklaşım oldukça öbeklenebilirdir, zira potansiyel alt grafiklerin sayısı – ve dolayısıyla zorlayıcı sorular – grafik boyutuyla doğrusal olmayan bir şekilde artarak verimli, büyük ölçüli veri sentezine olanak tanır。

Üretilen Seviye 3 görevlerimizin özelliklerini açıklamak amacıyla aşağıda iki örnek sunulmuştur. Bu sorular, metodolojimizin özünü teşkil eder: birden fazla, karmaşık bir şekilde bağlantılı varlıklar ve kasten belirsizleştirilmiş bilgiler içerir; örneğin belirsiz zaman referansları ("5. yüzyıl ortaları civarında", "21. yüzyıl başları") ve spesifik olmayan tanımlayıcılar ("onde gelen bir Güney Amerika başkenti", "saygın bir sanat kurumu") gibi. Yapısal karmaşıklık ve bilisel belirsizliğin bu birleşimi, giderilmesi son derece zor olan yüksek derecede başlangıç belirsizliği yaratmaktadır. Nitekim, manuel değerlendirmelerimiz, arama için net başlangıç noktalardan yoksun olmaları ve kapsamlı, doğrusal olmayan bir keşif gerektirmeleri nedeniyle, bu sorunların tipik zaman kısıtlamaları altında (örneğin, iki saat içinde) insan araştırmacılar için çözümsüz olduğunu doğrulamaktadır. Soru-cevap üretim sürecimiz hakkında daha fazla ayrıntı Ek A.2'de mevcuttur.

#### Oluşturulan Sorulardan Örnekler

**Soru:** *5. yüzyılın ortalarında vefat eden geç antik döneme ait bir yazar tarafından bestelenmiş erken dönem Hristiyan şiirsel bir ilahi bulunmaktadır. Bu yazarın ölüm yılı, modern çağdan birkaç yüzyıl öncesine ait çevresel koşulları yeniden yapılanlayan bilimsel bir kronolojinin son yılına denk gelmektedir. Bu kronolojinin adı nedir?*

**Cevap:** *Tahmini Ağaç Halkası Kronolojisi: 300-450 A.D.*

**Soru:** *Tanınmış bir Güney Amerika başkentiyle yakından ilişkili olan bir müzik eseri, 21. yüzyılın başlarında önemli bir yerel sivil onurla tanınan dikkate değer bir figür tarafından yazılın sözlerle sahiptir. Bu kompozisyonun melodisi, Batı Kolombiya'daki saygın bir sanat kurumunda resmi eğitim almış bir müzisyen tarafından bestelenmiştir. Bu müzik eserinin adı nedir?*

**Cevap:** *Rue de Rivoli*

### 3.2 Uzman LRM Yörüngeinden Akıl Yürütmenin Yeniden Yapılandırılması

Karmaşık soru-cevap çiftlerini sentezledikten sonraki bir sonraki zorluk, soğuk başlangıç denetimi için ilgili çözüm yörüngeini üretmektedir. QwQ-32B (Qwen Ekibi, 2025) gibi güçlü açık kaynak LRM'ler bazı doğru yörünge sağlayabile de, bunların tüm çıktısını doğrudanince ayar için kullanmak ve rımsızdır. İki kritik sorun tespit ettik:

- **Biçimsel Kirlilik:** Bu LRM'ler, akıl yürütmeinde güçlü ve genellikle ayrıntılı önceliklere sahiptir. Bu çıktılar üzerinde doğrudanince ayar yapmak, ajanın kendi keşif stratejilerini geliştirme ve görülmemiş sorumlara genelleme yeteneğini körelterek aşırı kuralçı olabilir.
- **Bağlam Aşırı Yüklenmesi:** Akıl yürütme zincirlerinin ayrıntılı yapısı, karmaşık web ajanı görevleri için pratik bir engel teşkil etmektedir. Dutzinlerce araç çağrıları içeren bir yörunge, bağlam limitlerini aşan bir geçmiş kolayca oluşturulabilir; bu durum, performansı düşürür ve akıl yürütme sürecini zorlaştırır.

The process is as follows: first, we prompt an expert open-source LRM to generate a complete solution trajectory, including its native thoughts. From this full trajectory, we selectively discard the LRM’s original, verbose thoughts, retaining only the successful action-observation sequence  $(a_0, o_0, a_1, o_1, \dots)$ . This trace represents the “what” and “how” of the solution path, but not the “why”.

Next, we reconstruct the missing “why”. For each step  $t$  in the action trace, we possess the history up to the previous step,  $\mathcal{H}_{t-1} = (\hat{\tau}_0, a_0, o_0, \dots, \hat{\tau}_{t-1}, a_{t-1}, o_{t-1})$ , along with the expert’s chosen action  $a_t$  and the subsequent observation  $o_t$ . We then prompt a separate, powerful instruction-following model,  $\pi^*$ , to generate a new thought  $\hat{\tau}_t$  that serves as a concise, logical justification for taking action  $a_t$ :

$$\hat{\tau}_t \sim \pi^*(\tau | \mathcal{H}_{t-1}, a_t, o_t). \quad (2)$$

By iteratively applying this for every step, we synthesize a complete, high-quality reasoning trajectory  $\hat{\mathcal{H}}_T = (\hat{\tau}_0, a_0, o_0, \dots, \hat{\tau}_T, a_T, o_T)$  where the reasoning is clean and goal-oriented. For this reconstruction, we use another LLM and enforce a “short-CoT” style. This is a critical design choice, ensuring the final reasoning chain is compact enough for long-horizon tasks. This method allows us to scalably generate supervision data that instills complex reasoning patterns without the negative side effects of direct imitation.

## 4 Reinforcement Learning with Cold Start

Our training methodology is a two-stage process. Inspired by recent advancements in post-training (Chu et al., 2025; Swamy et al., 2025; Ye et al., 2025) which highlight the efficacy of targeted fine-tuning before more complex learning, we first employ a modest RFT phase as a “cold start”. This initial phase aims to equip the model with fundamental tool-use capabilities and adherence to the long-horizon reasoning skeleton. Subsequently, we leverage RL to further refine the agent’s reasoning abilities, enhance its sample efficiency (Yue et al., 2025), and enable fuller utilization of our high-quality, complex training data.

### 4.1 Rejection Sampling Fine-Tuning

**Setup** Within a complete trajectory  $\mathcal{H}_T$ , the agent’s thoughts ( $\tau_i$ ) are enclosed by `<think>` and `</think>` tags. Actions ( $a_i$ ) are demarcated by `<tool_call>` and `</tool_call>` for function calls, or `<answer>` and `</answer>` for final responses. The environment’s observations ( $o_i$ ) resulting from tool calls are wrapped with `<tool_response>` and `</tool_response>` tags. Different segments are separated by these special tokens.

**Filtering** We apply a *three-stage* filtering process to the expert-generated trajectories. Firstly, to guarantee the correctness of the supervisory signal, we conduct rejection sampling that only trajectories culminating in a correct final answer are retained. Secondly, acknowledging that the expert models possess superior long-context processing capabilities compared to our policy model, we discard any trajectory exceeding 32k tokens in length. Thirdly, we filter for task complexity by retaining those trajectories with more than 5 tool calls, as intricate reasoning patterns and effective planning strategies typically manifest through a more extended sequence of decision-making steps.

**Training objective** The training objective is to specifically enhance the agent’s decision-making capability—that is, its ability to generate effective thoughts and actions. Consequently, the tokens corresponding to the environment’s observations ( $o_i$ ) are masked out from the loss calculation (Chen et al., 2023).

### 4.2 Duplicating Sampling Policy Optimization

Following the RFT cold-start phase, which equips the model with fundamental tool-use capabilities and adherence to a reasoning skeleton, we propose Duplicating Sampling Policy Optimization (DUPO) to further refine the reasoning abilities, enhance the sample efficiency (Yue et al., 2025) and ultimately elicit

Süreç şu şekildedir: İlk olarak, yerel düşünceleri de dahil olmak üzere eksiksiz bir çözüm yörüngesi oluşturma için uzman bir açık kaynak LRM’yi yönlendiririz. Bu tam yöründeden, LRM’nin orijinal, ayrıntılı düşüncelerini seçici olarak atarız; yalnızca başarılı eylem-gözlem dizisini  $(\hat{\tau}_0, a_0, o_0, \hat{\tau}_1, a_1, o_1, \dots)$  koruruz. Bu iz, çözüm yolunun “ne”sinı ve “nasıl”ını temsil eder, ancak “neden”ini değil.

Ardından, eksik “neden”i yeniden yapılandırırız. Eylem izindeki her adım  $t$  için, önceki adıma kadar olan geçmişe,  $\mathcal{H}_{t-1} = (\hat{\tau}_0, a_0, o_0, \dots, \hat{\tau}_{t-1}, a_{t-1}, o_{t-1})$  sahibiz; uzmanın seçtiği eylem  $a_t$  ve sonraki gözlem  $o_t$  ise bunlarla birlikte verilir. Daha sonra, ayrı ve güçlü bir talimat takip eden model olan  $\pi^*$ ’a,  $a_t$  eylemini gerçekleştirmek için özlü ve mantıksal bir gerekçe olarak hizmet eden yeni bir düşünce  $\hat{\tau}_t$  üretmesi için komut veririz:

$$\hat{\tau}_t \sim \pi^*(\tau | \mathcal{H}_{t-1}, a_t, o_t). \quad (2)$$

Bu işlemi her adım için yinelemeli olarak uygulayarak, muhakemenin net ve hedef odaklı olduğu eksiksiz, yüksek kaliteli bir muhakeme yörüngesi  $\hat{\mathcal{H}}_T = (\hat{\tau}_0, a_0, o_0, \dots, \hat{\tau}_T, a_T, o_T)$  sentezleriz. Bu yeniden yapılandırma için başka bir Büyük Dil Modeli (LLM) kullanırız ve “kısa-CoT” stilini zorunlu kılarız. Bu, uzun ufuklu görevler için nihai muhakeme zincirinin yeterince kompakt olmasını sağlayan kritik bir tasarım tercihidir. Bu yöntem, doğrudan taklidin olumsuz yan etkileri olmaksızın, karmaşık akıl yürütme kalıplarını aşışan denetim verilerini ölçeklenebilir bir şekilde oluşturmamızı sağlar.

## 4 Soğuk Başlangıç ile Pekiştirmeli Öğrenme

Eğitim metodolojimiz iki aşamalı bir süreçtir. Daha karmaşık öğrenmeden önce hedefe yönelik ince ayarın etkinliğini vurgulayan son post-training gelişmelerinden (Chu ve diğerleri, 2025; Swamy ve diğerleri, 2025; Ye ve diğerleri, 2025) ilham alarak, ilk olarak bir “soğuk başlangıç” olarak mütevazı bir RFT aşaması uyguluyoruz. Bu başlangıç aşaması, modeli temel araç kullanma yetenekleri ve uzun ufuklu akıl yürütme iskeletine bağlılıkla donatmayı amaçlamaktadır. Daha sonra, aracının akıl yürütme yeteneklerini daha da geliştirmek, örnek verimliliğini artırmak (Yue ve diğerleri, 2025) ve yüksek kaliteli, karmaşık eğitim verilerimizin daha tam kullanılmasını sağlamak için pekiştirmeli öğrenmeden (RL) yararlanıyoruz.

### 4.1 Reddetme Örneklemesi ile İnce Ayar

Kurulum Eksiksiz bir yörüngede içinde  $\mathcal{H}_T$ , aracının düşünceleri ( $\tau_i$ ) şunlarla çevrelenir: `<think>` ve `</think>` etiketleri. Eylemler ( $a_i$ ), aşağıdaki ile belirlenir: `<tool_call>` ve `</tool_call>` işlev çağrıları için veya `<answer>` ve `</answer>` nihai yanıtlar için. Araç çağrılarından kaynaklanan ortamın gözlemleri ( $o_i$ ) aşağıdaki ile sarmalanır: `<tool_response>` ve `</tool_response>` etiketleri. Farklı segmentler bu özel belirteçlerle ayrılır.

Filtreleme Uzman tarafından üretilen yörüngelere üç aşamalı birfiltreleme süreci uyguluyoruz. İlk olarak, denetleyici sinyalin doğruluğunu garanti etmek amacıyla, yalnızca doğru bir nihai yanıtlı sonuçlanan yörüngelerin korunduğu bir reddetme örneklemesi gerçekleştiriyoruz. İkinci olarak, uzman modellerin politika modelimize kıyasla üstün uzun bağlam işleme yeteneklerine sahip olduğunu göz önünde bulundurarak, 32 bin token uzunluğunu aşan tüm yörüngeleri eliyoruz. Üçüncü, görev karmaşıklığını filtreleyerek, 5’ten fazla araç çağrıları içeren yörüngeleri muhafaza ederiz; zira karmaşık akıl yürütme kalıpları ve etkili planlama stratejileri tipik olarak daha uzun bir karar verme adımları dizisi aracılığıyla ortaya çıkar.

**Eğitim hedefi** Eğitim hedefi, ajanın karar verme yeteneğini, yani etkili düşünceler ve eylemler üretme kabiliyetini özel olarak geliştirmektir. Sonuç olarak, ortamın gözlemlerine ( $o_i$ ) karşılık gelen belirteçler, kayıp hesaplamasından maskelenir (Chen et al., 2023).

### 4.2 Yinelenen Örnekleme Politikası Optimizasyonu

Modeli temel araç kullanma yetenekleri ve bir akıl yürütme iskeletine bağlılıkla donatan RFT soğuk başlatma aşamasının ardından, akıl yürütme yeteneklerini daha da geliştirmek, örneklem verimliliğini artırmak (Yue et al., 2025) ve nihayetinde

its intrinsic potential to discover and internalize sophisticated problem-solving strategies beyond direct imitation.

The main difference between RL for agents and conventional reasoning tasks is that rollout is a multi-turn process involving interaction with the environment (tool responses) (Sun et al., 2024). However, the interaction with the environment causes the rollout speed of agent RL to be much slower compared to standard RL. DAPO (Yu et al., 2025) employs dynamic sampling to filter out rollouts that are entirely correct or incorrect, subsequently filling the batch to its target size with new QAs. While this is effective for data curation, it may necessitate sequential rollouts for different cases within the same batch. This sequential processing further exacerbates the slow training speeds characteristic of agentic RL.

To solve this issue, we first filter out overly simple cases (those with all 8 rollouts correct) before training. During training, instead of using padding to expand the batch, we duplicate samples from the same batch that have a non-zero standard deviation. Compared to DAPO’s dynamic sampling, this approach achieves approximately 2–3 times speedup. Similar to SFT, it is also necessary to mask observations when calculating the policy loss (Jin et al., 2025). We follow GPRO (Shao et al., 2024) to estimate the advantage in a group-relative manner. We also utilize the token-level policy gradient loss and higher clip techniques in DAPO. The training objective of DUPO is defined as follows:

$$\begin{aligned} \mathcal{J}(\theta) = & \mathbb{E}_{(q,y) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | \text{context})} \\ & \left[ \frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left( r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right] \quad (3) \\ \text{s.t. } & 0 < |\{o_i \mid \text{is\_equivalent}(y, o_i)\}| < G, \end{aligned}$$

where  $(q, y)$  is the question-answer pair,  $r_{i,t}(\theta)$  is the importance sampling ratio, and  $\hat{A}_{i,t}$  is an estimator of the advantage at time step  $t$ :

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid \text{context})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \text{context})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (4)$$

Notably, in Eq. 4,  $o_i$  represents the tokens generated by the model but not the whole trajectory. Meanwhile, *context* comprises the model generation and tool response. Cases with a standard deviation of 0 (i.e., all roll-out answers are either completely correct or completely incorrect) are removed. These slots in the batch were then filled by randomly duplicating other cases within the same batch whose standard deviation was not 0.

To avoid reward hacking (Amodei et al., 2016; OpenAI, 2025c), we adopt a rule-based reward that combines both format validation and answer validation:

$$R_i = 0.1 * R_i^{\text{format}} + 0.9 * R_i^{\text{answer}}. \quad (5)$$

Specifically, the format score verifies whether the rollout trajectory follows the predefined format, such as whether different content segments are correctly wrapped with tags like `<think>` and `<tool_call>`, and whether the sequence complies with the ReAct framework. The answer score uses an LLM as a judge to determine whether the final prediction is correct.

## 5 Experiments

### 5.1 Setup

**Models and Benchmarks** We perform RFT and RL training on Qwen-2.5-3B, Qwen-2.5-7B, Qwen-2.5-32B, Qwen-2.5-72B. We mainly evaluate our method on four challenging benchmarks:

- **BrowseComp-en** (Wei et al., 2025): one of the most challenging benchmarks introduced by OpenAI to evaluate the proficiency of AI agents in locating hard-to-find, often multi-faceted,

doğrudan taklidin ötesinde sofistike problem çözme stratejilerini keşfetme ve işselleştirme konusundaki içsel potansiyelini ortaya çıkarmak için Yineleme Örneklemesi Politikası Optimizasyonu (DUPO) öneriyoruz.

Ajanlar için RL ile geleneksel akıl yürütme görevleri arasındaki temel fark, rollout'un çevreyle (araç yanıtları) etkileşimi içeren çok turlu bir süreç olmasıdır (Sun et al., 2024). Ancak, çevre ile etkileşim, ajan Takviyeli Öğrenme'nin (RL) devreye alma hızının standart RL'ye kıyasla çok daha yavaş olmasına neden olur. DAPO (Yu vd., 2025), tamamen doğru veya yanlış olan devreye almalarıfiltrelemek için dinamik örneklemeye kullanır ve ardından yeni soru-cevap (QA) çiftleriyle toplu iş hedef boyutuna kadar doldurur. Bu yöntem veri kürasyonu için etkili olsa da, aynı toplu iş içindeki farklı durumlar için sıralı devreye almaları gerektirebilir. Bu sıralı işleme, ajan tabanlı Takviyeli Öğrenme'nin karakteristik yavaş eğitim hızlarını daha da kötüleştirmektedir.

Bu sorunu çözmek için, eğitimden önce aşırı basit durumları (yani, sekiz devreye almanın tümünün doğru olduğu durumları) filtreliyoruz. Eğitim sırasında, toplu işi genişletmek için dolgu kullanmak yerine, sıfır olmayan standart sapmaya sahip aynı toplu işten örnekleri çoğaltıyoruz. DAPO'nun dinamik örneklemesine kıyasla, bu yaklaşım yaklaşık 2-3 kat hızlanma sağluyor. SFT'ye benzer şekilde, politika kaybını hesaplarken gözlemleri maskelemek de gereklidir (Jin vd., 2025). Grup-göreli bir şekilde avantajı tahmin etmek için GPRO'yu (Shao vd., 2024) takip ediyoruz. Ayrıca DAPO'daki belirteç seviyesindeki ilke gradyan kaybını ve daha yüksek kırpmaya tekniklerini kullanıyoruz. DUPO'nun eğitim amacı aşağıdaki gibi tanımlanmıştır:

$$\begin{aligned} \mathcal{J}(\theta) = & \mathbb{E}_{(q,y) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | \text{context})} \\ & \left[ \frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left( r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right] \quad (3) \\ \text{s.t. } & 0 < |\{o_i \mid \text{is\_equivalent}(y, o_i)\}| < G, \end{aligned}$$

burada  $(q, y)$  soru-cevap çifti;  $r_{i,t}(\theta)$  önem örneklemeye oranı ve  $A^i_t$ ,  $t$  ise  $t$  zaman adımdındaki avantajın bir tahmincisidir:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid \text{context})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \text{context})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (4)$$

Özellikle, Denklem 4'te,  $o_i$  model tarafından üretilen jetonları temsil eder, ancak tüm yörüngeyi değil. Bu arada, bağlam model üretimini ve araç yanıtını kapsar. Standart sapması 0 olan durumlar (yani, tüm yayılma yanıtları ya tamamen doğru ya da tamamen yanlış olanlar) çıkarılır. Partideki bu boşluklar daha sonra, aynı parti içinde standart sapması 0 olmayan diğer durumların rastgele çoğaltılmasıyla dolduruldu.

Ödül manipülasyonundan kaçınmak amacıyla (Amodei vd., 2016; OpenAI, 2025c), hem format doğrulanmasını hem de yanıt doğrulamasını birleştiren kural tabanlı bir ödül sistemi benimsiyoruz:

$$R_i = 0.1 * R_i^{\text{format}} + 0.9 * R_i^{\text{answer}}. \quad (5)$$

Özellikle, format puanı, yayılma yörünge'sinin önceden tanımlanmış formata uyup uymadığını, örneğin farklı içerik segmentlerinin etiketlerle doğru şekilde sarılıp sarılmadığını doğrular. `<think>` ve `<tool_call>`, ve sırının ReAct çerçevesi ile uyumlu olup olmadığını kontrol eder. Yanıt puanı, nihai tahminin doğru olup olmadığını belirlemek için bir LLM'i yargıç olarak kullanır.

## 5 Deneyler

### 5.1 Kurulum

**Modeller ve Kıyaslamlar** Qwen-2.5-3B, Qwen-2.5-7B, Qwen-2.5-32B, Qwen-2.5-72B modelleri üzerinde RFT ve RL eğitimi gerçekleştirmekteyiz. Yöntemimizi esas olarak dört zorlu kıyaslama ile değerlendiriyoruz:

- **BrowseComp-en** (Wei vd., 2025): OpenAI tarafından, yapay zeka ajanlarının internet genelinde zor bulunan, genellikle çok yönlü,

information across the internet, which demands sophisticated browsing strategies and reasoning capabilities.

- **BrowseComp-zh** (Zhou et al., 2025): Similar to BrowseComp-en, but the QAs are in Chinese.
- **GAIA** (Mialon et al., 2023): A benchmark that requires multi-modality and tool-use abilities. We only use a subset of 103 cases from the text-only validation subset (Li et al., 2025c; Wu et al., 2025a).
- **XbenchDeepSearch** (Xbench-Team, 2025): A new, dynamic, professionally-aligned benchmark that focuses on evaluating AI agents’ tool usage capabilities, specifically in deep information retrieval and complex search tasks.

**Baselines** We compare our method with the following paradigms:

- **Direct Inference**: Models answer questions based on its internal knowledge. For the non-reasoning model, we choose Qwen-2.5-32B, Qwen-2.5-72B (Yang et al., 2024), GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025b), and for the reasoning models, we select QWQ-32B (Qwen Team, 2025), o4-mini (OpenAI, 2025c), and DeepSeek-R1 (Guo et al., 2025). We do not consider smaller models because their scores on BrowseComp are essentially zero.
- **Proprietary Browsing Agents**: We test OpenAI DeepResearch (OpenAI, 2025a), Grok-DeepResearch (x.ai, 2025), and Doubao with Deep Think and Search (Doubao, 2025); however, as not all of them are fully accessible via API, they were not tested across all benchmarks and experiments.
- **Open-source Agents**: We compare our method with recent open-source web/search agents, including Search-o1 (Li et al., 2025b), WebThinker (Li et al., 2025c), R1-Searcher (Song et al., 2025), and WebDancer (Wu et al., 2025a).

**Metric and Hyper-parameters** We default to pass@ $k$  evaluation (Chen et al., 2021) and report pass@1 using non-zero temperature, and temperature and top-p are set to 0.6 and 0.95. For accuracy, we use LLM as a judge (Liu et al., 2024; Wang et al., 2024). The pass@1 is computed as:

$$\text{pass}@1 = \frac{1}{n} \sum_{i=1}^n p_i, \quad (6)$$

where  $p_i$  denotes the correctness of the  $i$ -th response. For pass@ $k$  that  $k > 1$  we repeatedly generate for  $k$  times.

## 5.2 Main Results

Our main results, presented in Table 1, reveal several critical insights.

**The Inadequacy of Direct Inference for Complex Information Seeking** Relying solely on a model’s internal knowledge and reasoning capabilities is insufficient for solving complex information retrieval tasks. Across the board, all models, including strong proprietary model like GPT-4.1, exhibit poor performance on BrowseComp-en/zh, with accuracy scores often near zero. This starkly demonstrates that these tasks require dynamic interaction with an external information source—the web—to gather the necessary evidence. The inherent uncertainty and specificity of the questions far exceed the scope of pre-trained knowledge, underscoring the necessity of an agentic, tool-using framework.

**Superior Reasoning Models Show a Glimmer of Potential** Within the direct inference paradigm, a notable exception is the superior performance of leading reasoning models like DeepSeek-R1 and o4-mini compared to other base models. For instance, DeepSeek-R1 achieves a score of 26.3 on BrowseComp-zh, significantly higher than other models in its category. This suggests that their advanced intrinsic

bilgileri konumlandırma yetkinliğini değerlendirmek amacıyla sunulan en zorlu kıyaslamalardan biridir; bu da sofistike tarama stratejileri ve muhakeme yetenekleri gerektirmektedir.

- **BrowseComp-zh** (Zhou vd., 2025): BrowseComp-en'e benzer olmakla birlikte, soru-cevaplar (QA'lar) Çince'dir.
- **GAIA** (Mialon vd., 2023): Çok modluluk ve araç kullanım yetenekleri gerektiren bir kıyaslama. Yalnızca metin tabanlı doğrulama alt kümesinden (Li vd., 2025c; Wu vd., 2025a) 103 vakadan oluşan bir alt küme kullanmaktayız.
- **XbenchDeepSearch** (Xbench-Team, 2025): Yapay zeka ajanlarının araç kullanım yeteneklerini, özellikle derin bilgi erişimi ve karmaşık arama görevlerinde değerlendirmeye odaklanan yeni, dinamik ve profesyonel olarak uyumlu bir kıyaslama.

Temel Yaklaşımlar Yöntemimizi aşağıdaki paradigmalarla karşılaştırıyoruz:

- **Doğrudan Çıkarım** : Modeller, soruları kendi iç bilgilerine dayanarak yanıtlar. Akıl yürütme dışı modeller için Qwen-2.5-32B, Qwen-2.5-72B (Yang et al., 2024), GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025b) modellerini; akıl yürütme modelleri için ise QWQ-32B (Qwen Ekibi, 2025), o4-mini (OpenAI, 2025c) ve DeepSeek-R1 (Guo et al., 2025) modellerini seçiyoruz. Daha küçük modelleri dikkate almıyoruz çünkü BrowseComp üzerindeki puanları esasen sıfırdır.
- **Tescilli Tarama Ajanları** : OpenAI DeepResearch (OpenAI, 2025a), Grok-DeepResearch (x.ai, 2025) ve Deep Think and Search özellikli Doubao'yı (Doubao, 2025) test ettik; ancak, bunların hepsi API aracılığıyla tam olarak erişilebilir olmadığından, tüm kıyaslamalarda ve deneylerde test edilmediler.
- **Açık Kaynak Ajanları** : Yöntemimizi Search-o1 (Li vd., 2025b), WebThinker (Li vd., 2025c), R1-Searcher (Song vd., 2025) ve WebDancer (Wu vd., 2025a) dahil olmak üzere güncel açık kaynak web/arama ajanlarıyla karşılaştırıyoruz.

**Metrik ve Hiper-parametreler** Varsayılan olarak pass@  $k$  değerlendirmesini (Chen vd., 2021) kullanınız ve sıfır olmayan sıcaklık kullanarak pass@1'i rapor ederiz; sıcaklık ve top-p sırasıyla 0.6 ve 0.95 olarak ayarlanmıştır. Doğru-luk için LLM'yi bir hakem olarak kullanınız (Liu vd., 2024; Wang vd., 2024). pass@1 şu şekilde hesaplanır:

$$\text{pass}@1 = \frac{1}{n} \sum_{i=1}^n p_i, \quad (6)$$

burada  $p_i$ ,  $i$ -inci yanıtın doğruluğunu belirtir.  $k > 1$  olan pass@ $k$  için  $k$  kez tekrar üretim yaparız.

## 5.2 Ana Sonuçlar

Tablo 1'de sunulan ana sonuçlarımız, çeşitli kritik içgörüler ortaya koymaktadır.

**Karmaşık Bilgi Arama İçin Doğrudan Çıkarımın Yetersizliği** Bir modelin yalnızca dahili bilgisine ve akıl yürütme yeteneklerine güvenmek, karmaşık bilgi alma görevlerini çözmek için yetersizdir. Genel olarak, GPT-4.1 gibi güçlü tescilli modeller de dahil olmak üzere tüm modeller, BrowseComp-en/zh üzerinde zayıf performans sergilemeye olup, doğruluk puanları genellikle sıfıra yakındır. Bu durum, bu görevlerin gerekli kanıtları toplamak için harici bir bilgi kaynağı olan web ile dinamik etkileşimi gerektirdiğini açıkça göstermektedir. Soruların doğasında var olan belirsizlik ve özgüllük, önceden eğitilmiş bilginin kapsamını çok aşmakta, ajansal, araç kullanan bir çerçeveyenin gerekliliğinin altını çizmektedir.

**Üstün Akıl Yürütme Modelleri Potansiyel Işığı Gösteriyor** Doğrudan çıkarım paradigmada, DeepSeek-R1 ve o4-mini gibi önde gelen akıl yürütme modellerinin diğer temel modellere kıyasla üstün performansı dikkate değer bir istisnadır. Örneğin, DeepSeek-R1, BrowseComp-zh üzerinde 26.3'lük bir puan elde ederek kendi kategorisindeki diğer modellerden önemli ölçüde daha yüksek bir performans sergilemektedir. Bu, onların gelişmiş içsel

Table 1: Main results on four challenging benchmarks. <sup>‡</sup> indicates that these proprietary methods are manually evaluated through their websites (some are reported in the corresponding benchmark papers). - means that we do not have the results due to cost constraints.

Backbone	Paradigm	BrowseComp-en	BrowseComp-zh	Xbench-DeepSearch	GAIA
<i>Direct Inference</i>					
Qwen-2.5-32B	Direct	0.6	3.9	8.7	13.6
Qwen-2.5-72B	Direct	0.6	7.0	12.7	14.6
GPT-4o	Direct	0.6	6.2	18.0	17.5
GPT-4.1	Direct	1.5	14.4	17.0	22.3
QwQ-32B	Direct	0.5	10.0	10.7	22.3
o4-mini	Direct	<b>6.1</b>	15.2	22.3	<b>33.3</b>
DeepSeek-R1	Direct	2.0	<b>26.3</b>	<b>32.7</b>	16.5
<i>Proprietary Agents</i>					
Grok-3 <sup>‡</sup>	Browsing	-	12.9	50+	-
Doubao <sup>‡</sup>	Browsing	-	26.0	50+	-
GPT-4o <sup>‡</sup>	Browsing	1.9	-	-	-
DeepResearch <sup>‡</sup>	Browsing	<b>51.5</b>	<b>42.9</b>	-	<b>67.4</b>
<i>Open-source Agents</i>					
R1-Searcher-7B	ReAct	0.4	0.6	4.0	20.4
Qwen-2.5-32B	Search-o1	0.1	2.4	3.7	28.2
WebDancer-32B	ReAct	2.5	14.1	38.7	40.7
QwQ-32B	Search-o1	2.8	17.9	25.0	39.8
WebThinker-RL	ReAct	2.8	7.3	24.0	48.5
WebDancer-QwQ	ReAct	3.8	18.0	39.0	51.5
WebSailor-3B	ReAct	3.3	9.7	27.7	33.0
WebSailor-7B	ReAct	6.7	14.2	34.3	37.9
WebSailor-32B	ReAct	10.5	25.5	53.3	53.2
WebSailor-72B	ReAct	<b>12.0</b>	<b>30.1</b>	<b>55.0</b>	<b>55.4</b>

reasoning capabilities allow them to better decompose complex questions and reduce uncertainty to some extent, even without external tools.

**WebSailor Establishes a New State-of-the-Art for Open-Source Agents** WebSailor sets a new state-of-the-art for open-source agents, with its advantage being most pronounced on the exceptionally challenging BrowseComp-en and BrowseComp-zh benchmarks. This result validates our core hypothesis: training on data synthesized to embody complex, hard-to-reduce uncertainty endows an agent with robust and generalizable reasoning strategies. The efficacy of our approach is strikingly demonstrated by WebSailor-3B and WebSailor-7B. Despite their modest size, WebSailor-7B achieves an accuracy of 6.7 on BrowseComp-en, decisively outperforming agents built on much larger 32B models, such as WebDancer-32B (2.5) and WebThinker-RL (2.8). This underscores that the performance gains are driven by our novel training paradigm—sophisticated data synthesis and targeted reinforcement learning—rather than being a mere artifact of model scale. While WebSailor performs strongly across all benchmarks, its margin on GAIA is more modest. Our manual inspection reveals this is because a significant portion of GAIA tasks requires mathematical and computational abilities, for which WebSailor was not specifically optimized. However, its accuracy on the purely information-retrieval subsets of GAIA remains exceptionally high, reaffirming its specialized expertise.

Tablo 1: Dört zorlu karşılaştırmalı değerlendirme üzerindeki ana sonuçlar. <sup>‡</sup> bu tescilli yöntemlerin web siteleri aracılığıyla manuel olarak değerlendirildiğini (bazları ilgili karşılaştırmalı değerlendirme makalelerinde bildirilmiştir) gösterir. - maliyet kısıtlamaları nedeniyle sonuçlara sahip olmadığımız anlamına gelir.

Temel Model	Paradigm	BrowseComp-en	BrowseComp-zh	Xbench-DeepSearch	GAIA
<i>Doğrudan Çıkarım</i>					
Qwen-2.5-32B	Doğrudan	0.6	3.9	8.7	13.6
Qwen-2.5-72B	Doğrudan	0.6	7.0	12.7	14.6
GPT-4o	Doğrudan	0.6	6.2	18.0	17.5
GPT-4.1	Doğrudan	1.5	14.4	17.0	22.3
QwQ-32B	Doğrudan	0.5	10.0	10.7	22.3
o4-mini	Doğrudan	<b>6.1</b>	15.2	22.3	<b>33.3</b>
DeepSeek-R1	Doğrudan	2.0	<b>26.3</b>	<b>32.7</b>	16.5
<i>Tescilli Ajanlar</i>					
Grok-3 <sup>‡</sup>	Tarama	-	12.9	50+	-
Doubao <sup>‡</sup>	Tarama	-	26.0	50+	-
GPT-4o <sup>‡</sup>	Tarama	1.9	-	-	-
DeepResearch <sup>‡</sup>	Tarama	<b>51.5</b>	<b>42.9</b>	-	<b>67.4</b>
<i>Açık Kaynak Ajanlar</i>					
R1-Searcher-7B	ReAct	0.4	0.6	4.0	20.4
Qwen-2.5-32B	Search-o1	0.1	2.4	3.7	28.2
WebDancer-32B	ReAct	2.5	14.1	38.7	40.7
QwQ-32B	Search-o1	2.8	17.9	25.0	39.8
WebThinker-RL	ReAct	2.8	7.3	24.0	48.5
WebDancer-QwQ	ReAct	3.8	18.0	39.0	51.5
WebSailor-3B	ReAct	3.3	9.7	27.7	33.0
WebSailor-7B	ReAct	6.7	14.2	34.3	37.9
WebSailor-32B	ReAct	10.5	25.5	53.3	53.2
WebSailor-72B	ReAct	<b>12.0</b>	<b>30.1</b>	<b>55.0</b>	<b>55.4</b>

akıl yürütme yetenekleri, karmaşık soruları daha iyi ayırtırmalarına ve harici araçlar olmaksızın dahi belirsizliği bir ölçüde azaltmalarına imkân tanır.

**WebSailor Açık Kaynaklı Ajanlar İçin Yeni Bir En Son Teknoloji Oluşturuyor** WebSailor, açık kaynaklı ajanlar için yeni bir en son teknoloji standarı belirlemekte olup, avantajı özellikle zorlu BrowseComp-en ve BrowseComp-zh kıyaslamalarında en belirgin şekilde ortaya çıkmaktadır. Bu sonuç, temel hipotezimizi doğrulamaktadır: karmaşık, azaltılması güç belirsizliği barındırmak üzere sentezenmiş veriler üzerinde yapılan eğitim, bir ajanı sağlam ve genellenebilir akıl yürütme stratejileriyle donatır. Yaklaşımımızın etkinliği, WebSailor-3B ve WebSailor-7B tarafından çarpıcı bir şekilde gösterilmiştir. Mütevazı boyutlarına rağmen, WebSailor-7B BrowseComp-en'de 6.7'lik bir doğruluk oranı elde ederek, WebDancer-32B (2.5) ve WebThinker-RL (2.8) gibi çok daha büyük 32B modellere dayalı ajanları kesin bir şekilde geride bırakmaktadır. Bu durum, performans artışlarının yalnızca model ölçüğünün bir sonucu olmaktan ziyade, yeni eğitim paradigmamız—sofistike veri sentezi ve hedeflenmiş pekiştirmeli öğrenme—tarafından yönlendirildiğini vurgulamaktadır. WebSailor, tüm kıyaslamalarda güçlü bir performans sergilemesine rağmen, GAIA üzerindeki farkı daha mütevazıdır. Manuel incelememiz, GAIA görevlerinin önemli bir kısmının matematiksel ve hesaplama yetenekleri gerektirmesinden kaynaklandığını ortaya koymaktadır; WebSailor bu yetenekler için özel olarak optimize edilmemiştir. Ancak, GAIA'nın salt bilgi-erişim alt kümelerindeki doğruluğu istisnai derecede yüksek kalmakta ve uzmanlaşmış yetkinliğini yeniden teyit etmektedir.

**Achieving Parity with Proprietary Systems** Perhaps the most significant finding is that WebSailor closes the gap between open-source and leading proprietary systems. On BrowseComp-zh, WebSailor-72B achieves performance on par with Doubao, a top-tier proprietary agent. While the SOTA system DeepResearch still holds a lead, WebSailor’s performance represents a major milestone, demonstrating that with sophisticated data synthesis and targeted training strategies like DUPO, open-source models can be elevated to a level of capability previously exclusive to closed, proprietary systems.

### 5.3 Analysis

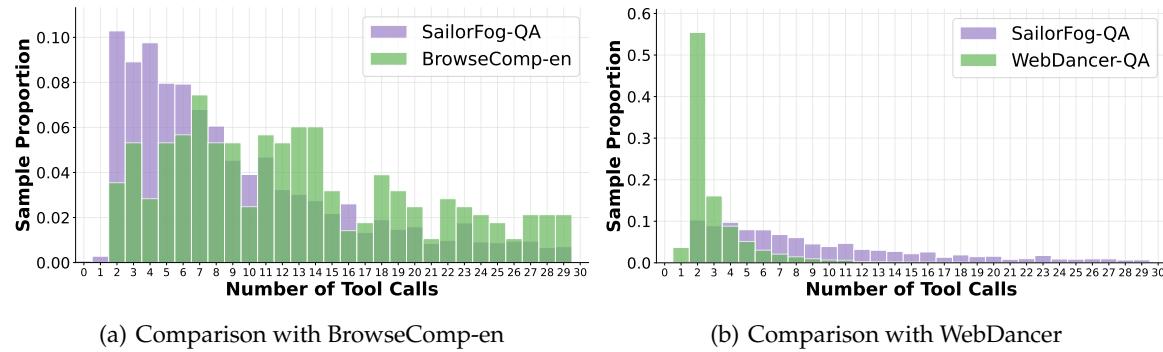


Figure 3: A comparison of the number of tool calls in our training set with those in the training sets of WebDancer and BrowseComp-en.

**Complexity of SailorFog-QA** Figure 3 provides a quantitative analysis of task complexity by plotting the distribution of tool call counts for our expert-generated training data against both the BrowseComp-en benchmark and the WebDancer training set. We use the number of tool calls as a proxy for problem difficulty. This analysis is based on unfiltered but correct trajectories from rejection sampling. The WebDancer dataset is heavily skewed towards simplicity, with over 50% of its trajectories requiring only two tool calls and virtually none exceeding ten. In sharp contrast, our synthesized data exhibits a long-tail distribution, with a significant concentration of samples requiring more than five tool calls and extending to trajectories with over twenty interactions. Crucially, this distribution closely mirrors the complexity profile of the BrowseComp-en benchmark itself. It is important to note that the figure displays our data before our final filtering stage, where we retain only trajectories with more than five tool calls. This strategic data construction ensures that our model is trained on problems that are not only complex but also structurally representative of the hard reasoning tasks, thereby equipping it with the robust, multi-step reasoning capabilities necessary for success.

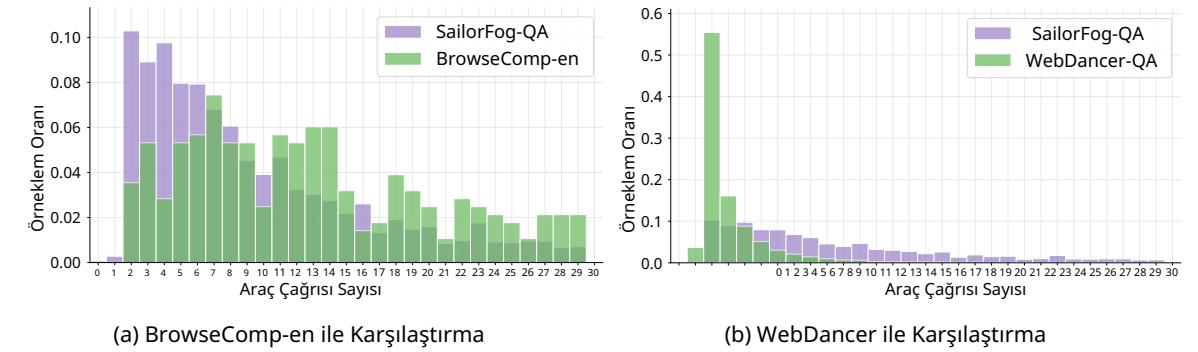
**Pass rate of SailorFog-QA** To further understand the difficulty of our synthetic data, Table 2 presents the pass@1 accuracy of SailorFog-QA, the WebDancer training set, and BrowseComp-en before filtering. DeepSeek-R1 and o4-mini are equipped with browsing tools and ReAct framework. We observe that, before filtering, our data is significantly more difficult than the WebDancer training set. Although the difficulty is lower than BrowseComp-en, it is worth noting that BrowseComp-en filters out simple cases (Wei et al., 2025). Upon manual inspection, we find that the low accuracy in our data is partly due to its inherent difficulty, but also because there may not always be a unique answer. Ambiguity in the information can result in multiple intersections of conditions that do not yield a single definitive answer—this is similar to the situation in BrowseComp-en. However, we

Table 2: The pass@1 accuracy of the SailorFog-QA, the WebDancer training set, and BrowseComp-en under the ReAct framework.

Backbone	SailorFog-QA	WebDancer-QA	BrowseComp-en
o4-mini	47.3	90.2	26.3
DeepSeek-R1	38.9	84.4	9.5

**Tescilli Sistemlerle Eşitliğin Sağlanması** Belki de en önemli bulgu, WebSailor’ın açık kaynak ve önde gelen tescilli sistemler arasındaki farkı kapatmasıdır. BrowseComp-zh üzerinde, WebSailor-72B, üst düzey tescilli bir ajan olan Doubao ile eş düzeyde performans göstermektedir. SOTA sistemi DeepResearch hala liderliğini sürdürse de, WebSailor’ın performansı önemli bir kilometre taşı temsil etmektedir; bu da sofistike veri sentezi ve DUPO gibi hedeflenmiş eğitim stratejileriyle açık kaynak modellerinin, daha önce kapalı, tescilli sistemlere özgü olan bir yetenek seviyesine yükseltilebileceğini göstermektedir.

### 5.3 Analiz



Şekil 3: Eğitim setimizdeki araç çağrıları sayılarının WebDancer ve BrowseComp-en eğitim setlerindeki karşılaştırması.

**SailorFog-QA’nın Karmaşıklığı** Şekil 3, uzmanlar tarafından oluşturulan eğitim verilerimizdeki araç çağrıları sayılarının dağılımını hem BrowseComp-en kıyaslama ölçüdü hem de WebDancer eğitim seti ile karşılaştırarak görev karmaşıklığının nicel bir analizini sunar. Problem zorluğunun bir göstergesi olarak araç çağrıları sayısını kullanıyoruz. Bu analiz, ret örneklemesinden elde edilen filtrelenmemiş ancak doğru yönlendirme dayanmaktadır. WebDancer veri seti, %50’den fazlasının yalnızca iki araç çağrıları gerektirmesi ve neredeyse hiçbirinin onu aşmamasıyla büyük ölçüde basitleştirme yönelik. Buna keskin bir tezat oluşturan sentezlenmiş verilerimiz, beşten fazla araç çağrıları gerektiren örneklerin önemli bir yoğunlaşması ve yirmiden fazla etkileşim içeren yönlendirme kadar uzanan uzun kuyruklu bir dağılım sergilemektedir. En önemlisi, bu dağılım BrowseComp-en kıyaslama ölçütünün karmaşıklık profilini yakından yansımaktadır. Şunu belirtmek önemlidir ki, şekil, beşten fazla araç çağrıları içeren yönlendirmeleri koruduğumuz son filtreleme aşamamızdan önceki verilerimizi göstermektedir. Bu stratejik veri inşası, modelimizin yalnızca karmaşık olmayıp, aynı zamanda zorlu akıl yürütme görevlerinin yapısal temsilci olan sorunlar üzerinde eğitilmesini sağlayarak, başarı için gerekli sağlam, çok adımlı akıl yürütme yetenekleriyle donatılmasını temin eder.

#### SailorFog-QA’nın geçiş oranı

Sentetik verilerimizin zorluğunu daha iyi anlamak için, Tablo 2, filtrelemeden önce SailorFog-QA’nın pass@1 doğruluğunu sunmaktadır.

DeepSeek-R1 ve o4-mini, gezinme araçları ve ReAct çerçevesi ile donatılmıştır. Filtrelemeden önce verilerimizin WebDancer eğitim setinden önemli ölçüde daha zor olduğunu gözlemezmekteyiz. Zorluk BrowseComp-en’den daha düşük olmasına rağmen, şunu belirtmekte fayda var ki

BrowseComp-en, basit durumları eler (Wei vd., 2025). Manuel inceleme sonucunda, verilerimizdeki düşük doğruluğun kısmen içeriği zorluktan, ancak aynı zamanda her zaman tekil bir cevabın bulunmamasından kaynaklandığını tespit ettik. Bilgideki belirsizlik, tek ve kesin bir yanıt vermeyen koşulların birden fazla kesimine yol açabilir; bu durum BrowseComp-en’deki senaryoya benzerdir. Ancak, biz

Tablo 2: SailorFog-QA, WebDancer eğitim seti ve BrowseComp-en'in ReAct çerçevesi altındaki pass@1 doğruluğu.

Temel Model	SailorFog-QA	WebDancer-QA	BrowseComp-en
o4-mini	47.3	90.2	26.3
DeepSeek-R1	38.9	84.4	9.5

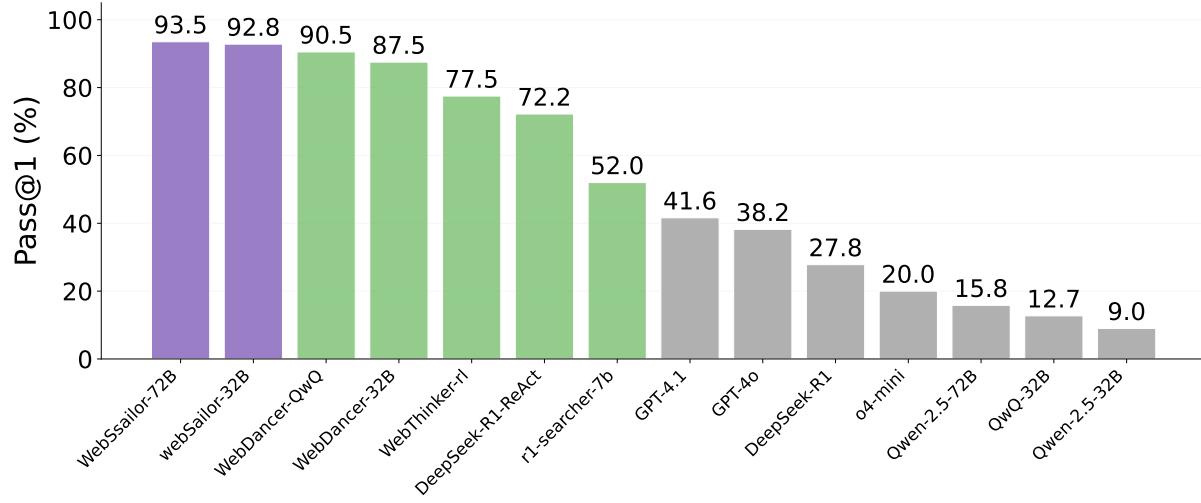


Figure 4: Performance on the SimpleQA benchmark.

can ensure the correctness of the conditions relative to the answer, i.e., the answer always satisfies the constraints specified in the question.

**Compatibility with simple tasks** WebSailor is trained exclusively on high-difficulty data, while BrowsecComp-en/zh, GAIA, and Xbench can all be considered as level-2 or level-3 tasks according to our definition. To verify whether WebSailor still performs strongly on simpler level-1 tasks, we evaluate its performance on a subset of SimpleQA benchmark (Wei et al., 2024). The complete SimpleQA dataset contains 4,326 QA pairs. Since testing on the entire set would be too time-consuming, we randomly sample 200 QA pairs for evaluation. This benchmark is characterized by high correctness and fact-based questions with simple conditions, and it is challenging for frontier LLMs to answer directly. The results, as shown in Figure 4, indicate that almost all agent-based methods outperform direct answering. WebSailor surpasses all other methods, demonstrating its compatibility and effectiveness even on simpler tasks.

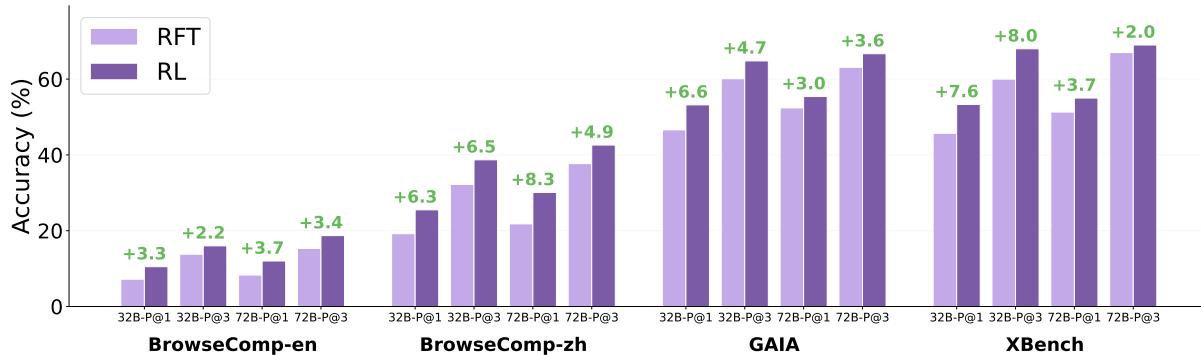
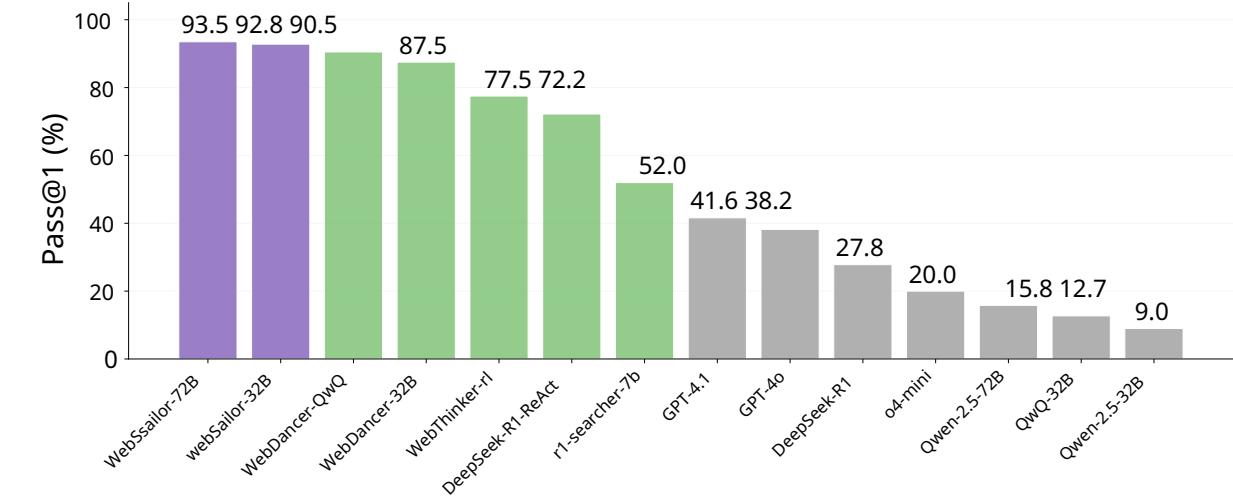


Figure 5: Detailed evaluation results using Pass@1, Pass@3.

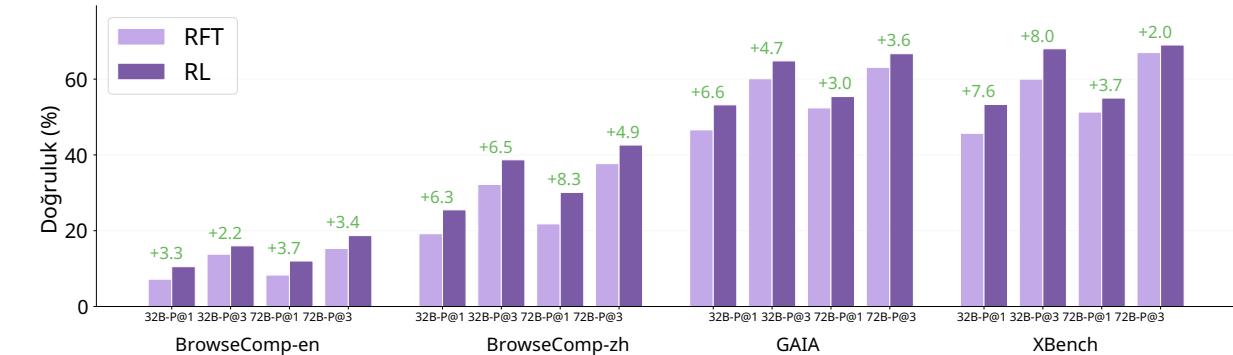
**Pass@1 vs Pass@3** We analyze the impact of our RL training by comparing the Pass@1 and Pass@3 performance of WebSailor before and after the RL stage (Fig. 5). The results reveal that RL brings notable improvements across all benchmarks, with the most significant gains observed on the highly difficult BrowsecComp-en/zh tasks. This disparity is telling: the extreme complexity of BrowsecComp requires agents to generate exceptionally long and intricate trajectories, making stable, repeatable success challenging (Sun et al., 2025). This instability is evident in the wide initial gap between Pass@1 and Pass@3 scores for BrowsecComp. RL training directly addresses this issue by reinforcing successful



Şekil 4: SimpleQA kıyaslama testi üzerindeki performans.

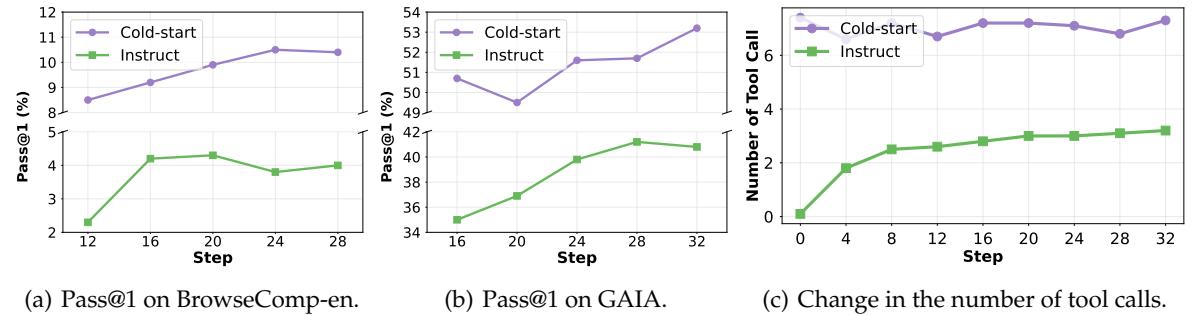
cevaba göre koşulların doğruluğunu temin edebiliriz, yani cevap her zaman soruda belirtilen kısıtlamaları karşılar.

**Basit görevlerle uyumluluk** WebSailor, yalnızca yüksek zorluktaki veriler üzerinde eğitilmiştir; oysa BrowsecComp-en/zh, GAIA ve Xbench'in tümü, tanımımıza göre seviye-2 veya seviye-3 görevleri olarak kabul edilebilir. WebSailor'ın daha basit seviye-1 görevlerde hâlâ güçlü performans sergileyip sergilemediğini doğrulamak amacıyla, SimpleQA karşılaştırma setinin bir alt kümesi üzerinde performansını değerlendirdik (Wei vd., 2024). Eksiksiz SimpleQA veri seti 4.326 Soru-Cevap çifti içermektedir. Tüm set üzerinde test yapmak oldukça zaman alıcı olacağından, değerlendirme için rastgele 200 Soru-Cevap çifti örneklenmiştir. Bu karşılaştırma seti, yüksek doğruluklu ve basit koşullara sahip, gerçeklere dayalı sorularla karakterizedir ve öncü LLM'lerin doğrudan yanıtlanması için zorlayıcıdır. Şekil 4'te gösterilen sonuçlar, neredeyse tüm ajan tabanlı yöntemlerin doğrudan yanıtlama yöntemlerinden daha iyi performans sergilediğini göstermektedir. WebSailor, diğer tüm yöntemleri geride bırakarak daha basit görevlerde bile uyumluluğunu ve etkinliğini ortaya koymaktadır.



Şekil 5: Pass@1, Pass@3 kullanılarak elde edilen detaylı değerlendirme sonuçları.

**Pass@1 ve Pass@3 RL eğitiminin etkisini**, WebSailor'ın RL aşamasından önceki ve sonraki Pass@1 ve Pass@3 performanslarını karşılaştırarak analiz ediyoruz (Şekil 5). Sonuçlar, RL'nin tüm kıyaslama testlerinde belirgin iyileştirmeler sağladığını, en önemli kazanımların ise son derece zorlu BrowsecComp-en/zh görevlerinde gözlemlendiğini ortaya koymaktadır. Bu farklılık düşündürücüdür: BrowsecComp'un aşırı karmaşıklığı, ajanların istikrarlı ve tekrarlanabilir başarısı zorlaştıran son derece uzun ve karmaşık yörüngeler üretmesini gerektirir (Sun ve diğerleri, 2025). Bu istikrarsızlık, BrowsecComp için Pass@1 ve Pass@3 skorları arasındaki geniş başlangıç farkında açıkça görülmektedir. RL eğitimi, başarılı stratejileri pekiştirerek bu sorunu doğrudan ele almaktadır.



(a) Pass@1 on BrowseComp-en. (b) Pass@1 on GAIA. (c) Change in the number of tool calls.

Figure 6: Comparing direct RL training of Qwen-2.5-instruct-32B with RL training after an RFT cold start.

strategies and pruning ineffective ones, which significantly improves the model’s ability to converge on a coherent solution path. Consequently, the model’s stability is enhanced, leading to greater gains on harder benchmarks. Moreover, we observe that the improvement in Pass@1 is proportionally much larger than in Pass@3, indicating that RL substantially enhances sample efficiency (Yue et al., 2025), allowing the model to achieve near its full potential with just a single sample.

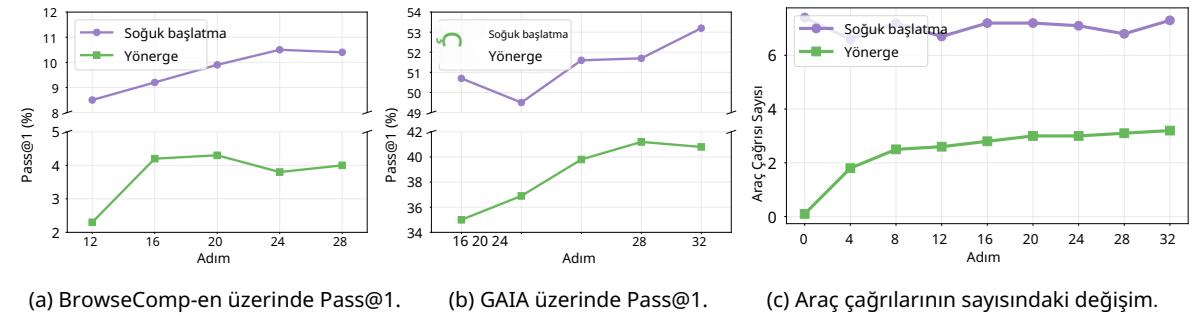
**With/without cold start** To evaluate the efficacy of our RFT cold-start strategy, we compare it against a direct RL training baseline, with results presented in Figure 6. We observe that while the direct RL approach exhibits a larger increase in Pass@1 accuracy, the final converged performance of the model that underwent an RFT cold start is significantly superior. This performance gap is also reflected in the models’ tool usage patterns. The tool call count for the cold-started model remains high and stable throughout RL training, whereas the tool call count for the direct RL model, despite a steady increase, remains substantially lower, indicating an inability to master long-horizon reasoning. Critically, the performance disparity between the cold-started WebSailor and the direct RL model is much wider on the BrowseComp-en. This suggests that without an RFT cold start, it is extremely difficult for a model to acquire the sophisticated reasoning patterns—often found only in powerful LRM—through self-exploration alone. The cold start is essential for bootstrapping the model with these complex strategies, which are necessary to solve exceptionally challenging tasks.

#### 5.4 Limitations and Future Work

First, our decision to filter training trajectories to under 32k tokens, while pragmatic, may cap the model’s ability to tackle even more complex problems. Our analysis of failed cases reveals that many errors stem from exceeding the context limit, and we observe that performance can degrade as inference length increases. Furthermore, WebSailor can exhibit a tendency for “over-thinking”, applying multi-step tool calls even to seemingly simple questions. However, this is not a clear-cut drawback; our qualitative analysis suggests that in many such instances, the agent is not aimlessly exploring but performing cross-verification, using different information sources to validate an initial finding. Finally, on the training front, our RL process is limited to 50 steps. This is primarily due to the inherent inefficiency of the synchronous RL framework; even with the optimizations from DUPO, the training speed remains a bottleneck. Future work will focus on migrating to an asynchronous training framework to improve efficiency and enable more extensive RL training.

## 6 Related Work

**Information seeking benchmarks** The landscape of information-seeking benchmarks has evolved from tasks with easily reducible uncertainty to those demanding complex, non-linear reasoning (Wu et al., 2025b). Early datasets such as NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and multi-hop variants like HotpotQA (Yang et al., 2018) and Musique (Trivedi et al., 2022) represent problems where



(a) BrowseComp-en üzerinde Pass@1. (b) GAIA üzerinde Pass@1. (c) Araç çağrılarının sayılarındaki değişim.

Şekil 6: Qwen-2.5-instruct-32B'nin doğrudan RL eğitimi ile bir RFT soğuk başlatma sonrası RL eğitiminin karşılaştırılması.

stratejileri ve etkisiz olanları eleyerek, modelin tutarlı bir çözüm yoluna yakınsama yeteneğini ölçüde geliştirir. Sonuç olarak, modelin istikrarı artırılır ve bu da daha zorlu kıyaslama testlerinde daha büyük kazanımlara yol açar. Dahası, Pass@1'deki iyileşmenin Pass@3'e kıyasla orantısız olarak çok daha büyük olduğunu gözlemliyoruz; bu da Pekiştirmeli Öğrenmenin (RL) örnek verimliliğini ölçüde artırdığını (Yue ve diğerleri, 2025) ve modelin tek bir örnekle tam potansiyeline yakın bir performans sergilemesini sağladığını gösteriyor.

**Soğuk başlatma ile/soğuk başlatma olmadan** RFT soğuk başlatma stratejimizin etkinliğini değerlendirmek amacıyla, sonuçları Şekil 6'da sunulan doğrudan RL eğitimi temel çizgisile karşılaştırma yapıyoruz. Doğrudan RL yaklaşımının Pass@1 doğruluğunda daha büyük bir artış sergilemesine rağmen, RFT soğuk başlatma uygulanan modelin nihai yakınlaşmış performansının ölçüde üstün olduğunu gözlemliyoruz. Bu performans farkı, modellerin araç kullanım kalıplarına da yansımaktadır. Soğuk başlatılan model için araç çağrıma sayısı, RL eğitimi boyunca yüksek ve istikrarlı kalırken, doğrudan RL modelinin araç çağrıma sayısı, istikrarlı bir artıya rağmen ölçüde düşük kalmaktadır; bu da uzun ufuklu akıl yürütmede ustalaşma yetersizliğini göstermektedir. Kritik olarak, soğuk başlatılan WebSailor ile doğrudan RL modeli arasındaki performans farkı BrowseComp-en üzerinde çok daha genişir. Bu durum, bir RFT soğuk başlatma olmaksızın, bir modelin –genellikle yalnızca güçlü LRM’lerde bulunan– karmaşık akıl yürütme kalıplarını yalnızca kendi kendine keşif yoluyla edinmesinin son derece zor olduğunu düşündürmektedir. Modeli, istisnai derecede zorlu görevleri çözmek için gerekli olan bu karmaşık stratejilerle başlatmak amacıyla soğuk başlangıç büyük önem taşımaktadır.

#### 5.4 Sınırlamalar ve Gelecek Çalışmaları

İlk olarak, eğitim yörüngelerini 32 bin belirtecin altında tutma kararımız, pragmatik olmakla birlikte, modelin daha da karmaşık sorunları ele alma yeteneğini kısıtlayabilir. Başarısız durum analizimiz, birçok hatanın bağlam sınırlarının aşılmasıından kaynaklandığını ortaya koymakta ve çıkarım uzunluğu arttıkça performansın düşebileceğini gözlemlemekteyiz. Ayrıca, WebSailor, görünüşte basit sorulara bile çok adımlı araç çağrıları uygulayarak “aşırı düşünme” eğilimi gösterebilir. Ancak, bu net bir dezavantaj değildir; niteliksel analizimiz, bu tür birçok durumda ajanın amaçsızca keşfetmek yerine, başlangıçtaki bir bulguya doğrulamak için farklı bilgi kaynaklarını kullanarak çapraz doğrulama yaptığından düşündürmektedir. Son olarak, eğitim aşamasında, RL sürecimiz 50 adımla sınırlıdır. Bu durum, temel olarak senkronize RL çerçevesinin doğasındaki verimsizlikten kaynaklanmaktadır; DUPO’nun optimizasyonlarına rağmen, eğitim hızı bir darboğaz olmaya devam etmektedir. Gelecekteki çalışmalar, verimliliği artırmak ve daha kapsamlı RL eğitimi sağlamak amacıyla eşzamansız bir eğitim çerçevesine geçmeye odaklanacaktır.

## 6 İlgili Çalışmalar

**Bilgi Arama Kıyaslamaları** Bilgi arama kıyaslamalarının kapsamı, kolayca azaltılabilir belirsizliğe sahip görevlerden, karmaşık, doğrusal olmayan akıl yürütme gerektiren görevlere doğru evrilmiştir (Wu vd., 2025b). NQ (Kwiatkowski vd., 2019), TriviaQA (Joshi vd., 2017) gibi erken dönem veri kümeleri ve HotpotQA (Yang vd., 2018) ile Musique (Trivedi vd., 2022) gibi çok adımlı varyantları, çözümlerin genellikle yapılandırılmış bir sorgu dizisi veya yalnızca bir modelin parametrik bilgisi aracılığıyla bulunabildiği problemleri temsil etmektedir.

solutions can often be found through a structured sequence of queries or even from a model’s parametric knowledge alone. More recent benchmarks have raised the complexity. GAIA (Mialon et al., 2023), while a generalist multimodal benchmark, introduces information-seeking challenges that approach complex multi-hop QA. Similarly, Xbench-DeepSearch (Xbench-Team, 2025) specifically targets agents’ deep search and tool-use capabilities through professionally annotated, dynamic tasks. At the apex of this evolution lie benchmarks like BrowseComp-en/zh (Wei et al., 2025; Zhou et al., 2025), which embody the Level 3 complexity central to our work. These tasks are characterized by intricately coupled entities and deliberate information obfuscation, creating high initial uncertainty that is exceptionally difficult to reduce. Success on BrowseComp requires the kind of sophisticated, non-linear exploration and synthesis that defines superhuman reasoning, making it the ideal proving ground for advanced web agents.

**Web agents** The development of autonomous web agents has witnessed significant progress from both proprietary and open-source communities (Zhang et al., 2025a). Proprietary systems like DeepResearch (OpenAI, 2025a), Doubao with Deep Think (Doubao, 2025), and Grok-3 (x.ai, 2025) have demonstrated superhuman performance in complex web navigation and information synthesis tasks, but their internal architectures and training methodologies remain opaque, impeding collaborative research. In contrast, open-source projects such as WebDancer (Wu et al., 2025a), WebThinker (Li et al., 2025c), and R1-Searcher (Song et al., 2025), adopting the ReAct framework (Yao et al., 2023), have made strides in simpler tasks yet face a substantial performance gap in benchmarks requiring sophisticated non-linear reasoning. In terms of training methodologies, the foundational principles of training dynamics outlined by (Xu et al., 2019; 2024; 2025) offer crucial insights, informing hyperparameter tuning and algorithmic design by highlighting the decisive impact of the training approach on models’ generalization ability. However, while many studies use Supervised Fine-Tuning (SFT) following the ReAct paradigm, pure SFT agents struggle with generalization in adaptive contexts (Zheng et al., 2025; Zhang et al., 2025b). Reinforcement learning-based methods Song et al. (2025); Zheng et al. (2025) hold promise for advanced search strategies via learned exploration policies but encounter challenges in training stability and sample efficiency.

## 7 Conclusion

In this work, we propose WebSailor. From the perspective of uncertainty reduction in information seeking, we analyze why previous open-source web agents have not reached the level of proprietary systems. Our contributions span from QA construction, comprehensive training data synthesis, RFT cold start, to improved efficiency in RL algorithms, leading to a full agentic post-training pipeline. WebSailor demonstrates strong performance on both simple and complex information seeking benchmarks, exhibiting reasoning and tool-use capabilities that surpass human levels.

We believe that the key to agentic post-training lies in further defining more complex tasks with higher uncertainty, as well as achieving more effective and efficient RL training. In the future, we will continue to explore how to further enhance agent capabilities based on open-source models, not only in the domain of information seeking, but also in pursuing the goal of general “superhuman” performance across more dimensions.

çözümler genellikle yapılandırılmış bir sorgu dizisi veya hatta yalnızca bir modelin parametrik bilgisi aracılığıyla bulunabilir. Daha yeni kıyaslamalar karmaşaklığını artırmıştır. GAIA (Mialon vd., 2023), genelci çok modlu bir kıyaslama olmasına rağmen, karmaşık çok adımlı Soru-Cevap (QA) sistemlerine yaklaşan bilgi arama zorlukları sunmaktadır. Benzer şekilde, Xbench-DeepSearch (Xbench-Team, 2025) özellikle profesyonelce açıklanmış, dinamik görevler aracılığıyla ajanların derinlemesine arama ve araç kullanma yeteneklerini hedeflemektedir. Bu evrimin zirvesinde, çalışmamızın merkezindeki Seviye 3 karmaşaklığını barındıran BrowseComp-en/zh (Wei vd., 2025; Zhou vd., 2025) gibi kıyaslama testleri yer almaktadır. Bu görevler, karmaşık bir şekilde birbirine bağlı varlıklar ve kasıtlı bilgi gizleme ile karakterize edilir; bu da başlangıçta azaltılması son derece güç olan yüksek bir belirsizlik ortamı yaratır. BrowseComp’ta başarı, süper insan akıl yürütmesini tanımlayan türden sofistike, doğrusal olmayan keşif ve sentez gerektirir; bu da onu gelişmiş web ajanları için ideal bir deneme alanı hâline getirir.

Web ajanları Otonom web ajanlarının geliştirilmesi, hem tescilli hem de açık kaynak topluluklarından önemli ilerlemeler kaydetmiştir (Zhang vd., 2025a). Deep- Research (OpenAI, 2025a), Deep Think ile Doubao (Doubao , 2025) ve Grok-3 (x.ai, 2025) gibi tescilli sistemler, karmaşık web navigasyonu ve bilgi sentezi görevlerinde süper insan performansı sergilemiş; ancak iç mimarileri ve eğitim metodolojileri şeffaf kalmaya devam ederek işbirliğine dayalı araştırmaları engellemektedir.

Buna karşılık, ReAct çerçevesini (Yao et al., 2023) benimseyen WebDancer (Wu et al., 2025a), WebThinker (Li et al., 2025c) ve R1-Searcher (Song et al., 2025) gibi açık kaynak projeleri, daha basit görevlerde ilerleme kaydetmiş olsa da, karmaşık doğrusal olmayan akıl yürütme gerektiren kıyaslamalarda önemli bir performans farkıyla karşılaşmaktadır. Eğitim metodolojileri açısından, (Xu et al., 2019; 2024; 2025) tarafından ana hatları çizilen eğitim dinamiklerinin temel ilkeleri, modellerin genelleme yeteneği üzerindeki eğitim yaklaşımının belirleyici etkisini vurgulayarak hiperparametre ayarı ve algoritmik tasarım hakkında kritik bilgiler sunmaktadır.

Ancak, birçok çalışma ReAct paradigmını takiben Denetimli İnce Ayar (SFT) kullanmasına rağmen, saf SFT ajanları adaptif bağlamlarda genelleme ile mücadele etmektedir (Zheng et al., 2025; Zhang et al., 2025b). Pekiştirmeli öğrenmeye dayalı yöntemler (Song et al., 2025; Zheng et al., 2025), öğrenilmiş keşif politikaları aracılığıyla gelişmiş arama stratejileri için umut vaat etse de, eğitim istikrarı ve örnek verimliliği konularında zorluklarla karşılaşmaktadır.

## 7 Sonuç

Bu çalışmada WebSailor’ı öneriyoruz. Bilgi arayışında belirsizliğin azaltılması perspektifinden, önceki açık kaynaklı web ajanlarının neden tescilli sistemlerin seviyesine ulaşmadığını analiz ediyoruz. Katkılarımız, QA oluşturmadan, kapsamlı eğitim verisi sentezine, RFT soğuk başlatmaya ve RL algoritmalarında artırılmış verimliliğe kadar uzanmakta, bu da eksiksiz bir ajan tabanlı son eğitim sürecine yol açmaktadır . WebSailor, hem basit hem de karmaşık bilgi arama karşılaştırmalı testlerinde güçlü bir performans sergileyerek, insan seviyelerini aşan akıl yürütme ve araç kullanma yetenekleri göstermektedir.

Ajan tabanlı son eğitimin anahtarının, daha yüksek belirsizliğe sahip karmaşık görevleri tanımlamak ve daha etkili ve verimli RL eğitimi sağlamak olduğuna inanıyoruz. Gelecekte, sadece bilgi arama alanında değil, aynı zamanda daha fazla boyutta genel “süper insan” performansı hedefini takip ederek, açık kaynaklı modellere dayalı ajan yeteneklerini nasıl daha da geliştireceğimizi keşfetmeye devam edeceğiz.

## A Experimental Details

### A.1 Tools

WebSailor uses two types of tools, search and visit:

- **Search** is used to access the Google search engine for information retrieval. The parameters of Search are the search queries. It allows searching multiple queries simultaneously and returns the top-10 results for each query. Each result contains a title, a snippet, and the corresponding URL.
- **Visit** is used to access specific web pages. The input consists of several web pages and their corresponding visit goals, with each page having a dedicated goal. First, Jina ([Jina.ai, 2025](#)) is used to retrieve the full content of the web page, and then a summary model extracts relevant information based on the goal. In this paper, we use Qwen-2.5-72B as the summary model.

### A.2 QA Construction

Our QA is constructed by sampling a subgraph from a graph. Each graph is generated by performing a random walk starting from a rare entity. The nodes in the graph represent entities, and the edges represent the relationships between entities. The general process for constructing the graph is as follows:

1. We use Wikidata's SPARQL service to obtain rare entities based on certain database rules.
2. The features of the initial node are obtained using the search and visit tools, and the initial node is set as the expansion node.
3. Some related entities are obtained based on the features of the expansion node, and then we get their features.
4. With a certain probability, we either set a new related entity as the next expansion node, or select a node from the previous nodes.
5. Repeat steps 3 and 4 until the number of edges in the graph reaches a predefined value.

### A.3 ReAct Trajectories

Our ReAct framework is implemented through Qwen-Agent<sup>1</sup>, and we limit the number of tool calls to no more than 30. A complete trajectory follows the format below:

#### Case Trajectory

```
<think> thinking process here </think>
<tool_call>
"name": "tool name here", "arguments": "parameter name here": parameter value here, "another
parameter name here": another parameter value here, ...
</tool_call>
<tool_response>
tool_response here
</tool_response>
(more thinking processes, tool calls and tool responses here)
<think> thinking process here </think>
<answer> answer here </answer>
```

## A Deneysel Detaylar

### A.1 Araçlar

WebSailor iki tür araç kullanır: arama ve ziyaret:

- Arama bilgi erişimi için Google arama motoruna erişimde kullanılır. Arama'nın parametreleri, arama sorgularıdır. Birden fazla sorgunun eş zamanlı olarak aranmasına olanak tanır ve her soru için ilk 10 sonucu döndürür. Her sonuç bir başlık, bir tanıtım metni ve ilgili URL içerir.
- Ziyaret belirli web sayfalarına erişimde kullanılır. Girdi, her birinin özel bir hedefi olan çeşitli web sayfaları ve bunlara karşılık gelen ziyaret hedeflerinden oluşur. İlk olarak, web sayfasının tam içeriğini almak için Jina ([Jina.ai, 2025](#)) kullanılır ve ardından bir özetleme modeli, hedefe dayalı ilgili bilgiyi çıkarır. Bu makalede, özetleme modeli olarak Qwen-2.5-72B'yi kullanmaktadır.

### A.2 Soru-Cevap Oluşturma

Soru-Cevap sistemimiz, bir grafikten bir alt grafinin örneklenmesiyle oluşturulur. Her grafik, nadir bir varlığıtan başlayarak rastgele bir gezinti gerçekleştirilerek üretilir. Grafikteki düğümler varlıklar, kenarlar ise varlıklar arasındaki ilişkileri temsil eder. Grafinin oluşturulmasına yönelik genel süreç aşağıdaki gibidir:

1. Belirli veritabanı kurallarına göre nadir varlıkları elde etmek için Wikidata'nın SPARQL hizmetini kullanıyoruz.
2. Başlangıç düğümünün özellikleri, arama ve ziyaret araçları kullanılarak elde edilir ve başlangıç düğümü genişleme düğümü olarak ayarlanır.
3. Genişleme düğümünün özelliklerine dayanarak bazı ilgili varlıklar elde edilir ve ardından bunların özellikleri alınır.
4. Belirli bir olasılıkla, ya yeni bir ilgili varlığı bir sonraki genişleme düğümü olarak belirleriz ya da önceki düğümlerden birini seçeriz.
5. Grafikteki kenar sayısı önceden tanımlanmış bir değere ulaşana kadar 3. ve 4. adımları tekrarlayın.

### A.3 ReAct Yörüneleri

ReAct çerçevesi Qwen-Agent<sup>1</sup> aracılığıyla uygulanmaktadır ve araç çağrılarının sayısını en fazla 30 ile sınırlıyoruz. Tam bir yörünge aşağıdaki formatı takip eder:

#### Vaka Yörünge

```
<think> düşünme süreci burada </think>
<tool_call>
"name": "araç adı burada", "arguments": "parametre adı burada": parametre değeri burada
, "başka bir parametre adı burada": başka bir parametre değeri burada, ...
</tool_call>
<tool_response>
araç_yanıtı burada
</tool_response>
(daha fazla düşünme süreci, araç çağrıları ve araç yanıtları burada)
<think> düşünme süreci burada </think>
<answer> yanıt burada </answer>
```

<sup>1</sup><https://github.com/QwenLM/Qwen-Agent/>

<sup>1</sup><https://github.com/QwenLM/Qwen-Agent/>

#### A.4 Training Details

We use Megatron (Shoeybi et al., 2019) for SFT and verl (Sheng et al., 2025) for RL training. For SFT, we use a batch size of 32, learning rate of 5e-6 with a minimum of 1e-10, warmup plus cosine decay schedule, and a weight decay of 0.1. For RL training, the rollout number in a group is 8, the temperature is 1.0,  $top_p = 1.0$ , the batch size is 128, the mini batch size is 32, and the learning rate is 1e-6.

#### A.5 Case Study

##### Case Trajectory in BrowseComp-en

**Question:** There is a software developer who, sometime after 2010 but before 2023, claimed to have designed and created a solar-powered fridge. They were known to live a rustic lifestyle sometime after 2012 and before 2023 and be located, in their words, "in a hole in the map". They have memories gleaned from a developer's conference in Edinburgh and were also known to enjoy caving. What was the make and model of the first computer purchased jointly with their father during the 1980s?

###### Step 1

<think>

The question requires identifying a software developer with a unique set of biographical details—someone who built a solar-powered fridge, lived a rustic lifestyle, described their location as “a hole in the map,” had memories from a developer’s conference in Edinburgh, and enjoyed caving. To make progress, I need to triangulate the individual’s identity by searching for overlapping evidence of these distinctive traits online...

</think>

<tool\_call>

```
{"name": "search", "arguments": {"query": ["software developer solar-powered fridge rustic lifestyle", "software developer hole in the map caving"]}}
```

</tool\_call>

<tool\_response>

A Google search for ‘software developer solar-powered fridge rustic lifestyle’ found 10 results:

- [Solar Fridge with Temperature Control and Monitoring System Using ...] (<https://www.ijert.org/solar-fridge-with-temperature-control-and-monitoring-system-using-bluetooth-module>)

Here we are designing a mini solar Bluetooth based refrigerator with temperature control with temperature monitoring system which is cheaper as well as eco- ...

- [Engineer, Embedded Software - EnPhase Energy - Climatebase] (<https://climatebase.org/job/61314385/engineer-embedded-software>) Date published: Jun 18, 2025

Enphase Energy is a global energy technology company and a leading provider of solar, battery, and electric vehicle charging products.

...

</tool\_response>

###### Step 2

<think>

The combination of unique biographical markers—such as a software developer with a history of building a solar-powered fridge, a self-described “hole in the map” rustic lifestyle, and a background in caving and developer conferences in Edinburgh—suggests that the right approach is to cross-reference these traits in more integrated search queries. By combining all these distinctive elements into multi-part searches...

#### A.4 Eğitim Detayları

SFT için Megatron (Shoeybi ve diğerleri, 2019) ve RL eğitimi için verl (Sheng ve diğerleri, 2025) kullanıyoruz . SFT için 32'lik bir toplu iş boyutu, minimum 1e-10 ile 5e-6'lık bir öğrenme oranı, isınma artı kosinus düşüşü programı ve 0.1'lik bir ağırlık düşüşü kullanıyoruz. RL eğitimi için bir gruptaki açılım sayısı 8, sıcaklık 1.0,  $top_p = 1.0$ , toplu iş boyutu 128, mini toplu iş boyutu 32 ve öğrenme oranı 1e-6'dır.

#### A.5 Vaka Çalışması

##### BrowseComp-en'deki Vaka Yörüngesi

**Soru:** 2010 sonrası ancak 2023 öncesi bir zamanda, güneş enerjili bir buzdolabı tasarladığını ve ürettiğini iddia eden bir yazılım geliştirici bulunmaktadır. Söz konusu kişinin, 2012 sonrası ancak 2023 öncesi bir dönemde rustik bir yaşam tarzı sürdürdü ve kendi ifadesiyle “haritadaki bir delikte” konumlandığı bilinmektedir. Edinburgh'daki bir geliştirici konferansından edindiği anıları vardı ve aynı zamanda mağaracılıktan hoşlandığı da biliniyordu. 1980'lerde başıyla birlikte ortaklaşa satın aldığı ilk bilgisayarın markası ve modeli neydi?

###### Adım 1

<think>

Soru, güneş enerjili bir buzdolabı inşa eden, rustik bir yaşam tarzı sürdürden, konumunu “haritadaki bir delik” olarak tanımlayan, Edinburgh'daki bir geliştirici konferansından anıları olan ve mağaracılıktan hoşlanan, benzersiz biyografik detaylara sahip bir yazılım geliştiricisini belirlemeyi gerektirmektedir. İlerlemek için, bu ayırt edici özelliklerin çevrimiçi olarak örtünen kanıtlarını arayarak bireyin kimliğini üçgenlemem gerekmektedir...

</think>

<tool\_call>

```
{"name": "search", "arguments": {"query": ["yazılım geliştirici güneş enerjili buzdolabı rustik yaşam tarzı", "yazılım geliştirici haritadaki delik mağaracılık"]}}
```

</tool\_call>

<tool\_response>

Yazılım geliştirici güneş enerjili buzdolabı rustik yaşam tarzı için yapılan bir Google araması 10 sonuç buldu:

- [Sıcaklık Kontrolü ve İzleme Sistemi Güneş Enerjili Buzdolabı ...] (<https://www.ijert.org/solar-fridge-with-temperature-control-and-monitoring-system-using-bluetooth-module>)

Burada, sıcaklık kontrollü ve sıcaklık izleme sisteme sahip mini bir güneş enerjili Bluetooth tabanlı buzdolabı tasarlıyoruz; bu hem daha ucuz hem de çevre dostu...

- [Mühendis, Gömülü Yazılım - EnPhase Energy - Climatebase] (<https://climatebase.org/job/61314385/engineer-embedded-software>) Yayınlanma tarihi: 18 Haziran 2025

Enphase Energy, küresel bir enerji teknolojisi şirketi ve onde gelen bir güneş enerjisi, bataraya ve elektrikli araç şarj ürünlerini sağlayıcıdır.

...

</tool\_response>

###### Adım 2

<think>

Güneş enerjili buzdolabı inşa etme geçmişi olan bir yazılım geliştirici, kendini ‘haritadaki delik’ olarak tanımlayan rustik bir yaşam tarzı ve Edinburgh'daki mağaracılık ve geliştirici konferansları geçmişi gibi benzersiz biyografik işaretlerin birleşimi, doğru yaklaşımın bu özellikleri daha entegre arama sorgularında çapraz referanslamak olduğunu düşündürmektedir. Tüm bu ayırt edici unsurları çok parçalı aramalarda birleştirerek...

```

</think>
<tool_call>
{"name": "search", "arguments": {"query": ["software developer rustic lifestyle caving conference Edinburgh", "solar-powered fridge caving conference Edinburgh rustic lifestyle"...]}}
</tool_call>
<tool_response>
A Google search for 'software developer rustic lifestyle caving conference Edinburgh' found 3 results:
1. [Developer conferences in Edinburgh 2025 / 2026 - dev.events](https://dev.events/EU/GB/Edinburgh)
Missing: rustic lifestyle caving
2. [Joy - The Red Hand Files](https://www.theredhandfiles.com/joy/)
Some have likened joy to be a flame enshrined behind a glass lantern – no matter how hard the wind blows, the flicker stays intact.
...
</tool_response>

```

### Step 3

<think>

The phrase “hole in the map” is highly distinctive and likely to be quoted or paraphrased by the individual in question or in descriptions about them. By searching for combinations involving “software developer solar-powered fridge” and “hole in the map,” I’m aiming to surface direct references—such as interviews, blog posts, or forum profiles—where the person has described their lifestyle and technical projects in their own words...

```

</think>
<tool_call>
{"name": "search", "arguments": {"query": ["software developer solar-powered fridge hole in the map", "software developer hole in the map rustic lifestyle"...]}}
</tool_call>
<tool_response>

```

A Google search for ‘software developer solar-powered fridge “hole in the map”’ found 5 results:

- [[PDF] Observing With The Green Bank Telescope - Science](https://www.gb.nrao.edu/glangsto/GBTog.pdf) Date published: Oct 11, 2011  
the largest “hole” in the map is approximately 0.3 x map radius. However it is useful for focal-plane arrays. Syntax: Daisy(location, map ...)
- [[PDF] Larissa Fergus.pdf - VU Research Repository](https://vuir.vu.edu.au/30098/1/Larissa%20Fergus.pdf)

I can see a tiny hole in the map, between contours, on the plain west of the mantelpiece. I lift my goggles and look over at the corresponding place on the

```

...
</tool_response>

```

### Step 4

<think>

The LWN.net article surfaced in the search results with a direct quote about living “in a hole in the map,” which matches the unique phrasing from the question and is strongly associated with a rustic lifestyle. This suggests the article may contain a first-person account or interview with the software developer in question, potentially revealing their identity and offering biographical details such as their interests in caving, solar-powered projects, and perhaps even recollections about early computing experiences...

```

</think>

```

```

</think>
<tool_call>
{"name": "search", "arguments": {"query": ["yazılım geliştirici rustik yaşam tarzı mağaracılık konferansı Edinburgh", "güneş enerjili buz dolabı mağaracılık konferansı Edinburgh rustik yaşam tarzı"...]}}
</tool_call>
<tool_response>
'Yazılım geliştirici rustik yaşam tarzı mağaracılık konferansı Edinburgh' için yapılan bir Google araması 3 sonuç buldu:
1. [Edinburgh 2025 / 2026 Geliştirici Konferansları - dev.events](https://dev.events/EU/GB/Edinburgh)
Eksik: rustik yaşam tarzı mağaracılık
2. [Neşe - The Red Hand Files](https://www.theredhandfiles.com/joy/)
Bazları neşeyi, cam bir fenerin içine yerleştirilmiş bir aleve benzetmiştir; rüzgar ne kadar sert eserse essin, titrek alev sağlam kalır.
...
</tool_response>

```

### Adım 3

<think>

“Haritadaki delik” ifadesi oldukça belirgin olup, söz konusu kişi tarafından veya o kişiyle ilgili açıklamalarda doğrudan alıntılanması veya yeniden ifade edilmesi muhtemeldir. “Yazılım geliştirici güneş enerjili buz dolabı” ve “haritadaki delik” içeren kombinasyonları arayarak, kişinin yaşam tarzını ve teknik projelerini kendi sözleriyle anlattığı doğrudan referansları (örneğin röportajlar, blog yazıları veya forum profilleri gibi) ortaya çıkarmayı hedefliyorum...

```

</think>
<tool_call>
{"name": "search", "arguments": {"query": ["yazılım geliştirici güneş enerjili buz dolabı haritadaki delik", "map", "yazılım geliştirici haritadaki delik", "rustic yaşam tarzı"...]}}
</tool_call>
<tool_response>

```

‘Yazılım geliştirici güneş enerjili buz dolabı “haritadaki delik”’ için yapılan Google araması 5 sonuç buldu:

- [[PDF] Green Bank Teleskopu İle Gözlem Yapmak - Bilim](https://www.gb.nrao.edu/glangsto/GBTog.pdf) Yayınlanma Tarihi: 11 Ekim 2011  
Haritadaki en büyük “delik” yaklaşık olarak harita yarıçapının 0.3 katıdır. Ancak odak düzleme dizileri için kullanışlıdır. Sözdizimi: Daisy(konum, harita ...)
- [[PDF] Larissa Fergus.pdf - VU Araştırma Deposu](https://vuir.vu.edu.au/30098/1/Larissa%20Fergus.pdf)  
Şöminenin batısındaki ovada, konturlar arasında, haritada küçük bir delik görebiliyorum. Gözlüğümü kaldırıp ... üzerindeki ilgili yere bakıyorum.

```

</tool_response>

```

### Adım 4

<think>

LWN.net makalesi, arama sonuçlarında “haritadaki bir delikte” yaşamakla ilgili doğrudan bir alıntıyla ortaya çıktı; bu ifade, sorudaki benzersiz söylemle eşleşmekte ve kırsal bir yaşam tarzıyla güçlü bir şekilde ilişkilendirilmektedir. Bu durum, makalenin söz konusu yazılım geliştiriciyle ilgili birinci açıdan bir anlatım veya röportaj içerebileceğini düşündürmektedir; bu da potansiyel olarak onların kimliğini ortaya çıkararak mağaracılık, güneş enerjili projeler gibi ilgi alanları ve hatta erken bilgisayar deneyimlerine dair anıları gibi biyografik ayrıntılar sunabilir

```
<tool_call>
{"name": "visit", "arguments": {"url": ["https://lwn.net/Articles/672352"], "goal": "Identify the software developer, their claim to have designed a solar-powered fridge, their rustic lifestyle mention, and any details about early computer purchases"}}
</tool_call>
```

```
<tool_response>
```

The useful information in <https://lwn.net/Articles/672352/> for user goal Identify the software developer, their claim to have designed a solar-powered fridge, their rustic lifestyle mention, and any details about early computer purchases as follows: ...

```
</tool_response>
```

### Step 5

```
<think>
```

The interview on LWN.net provides strong evidence that Joey Hess is the software developer in question, detailing his rustic lifestyle "in a hole in the map" and his early computing experiences. However, while there's mention of solar panels and off-grid living, there's no explicit confirmation yet that he claimed to have designed and created a solar-powered fridge—an important detail for fully verifying his identity as the subject of the question...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess solar-powered fridge", "Joey Hess solar powered fridge created", "Joey Hess designed solar-powered fridge", "Joey Hess fridge solar"]}}
```

```
</tool_call>
```

```
<tool_response>
```

A Google search for 'Joey Hess solar-powered fridge' found 9 results:

1. [Off-Grid, Solar-Powered, Zero-Battery Refrigerator](<https://www.notechmagazine.com/2019/09/off-grid-solar-powered-zero-battery-refrigerator.html>)

Date published: Sep 1, 2019

Joey Hess has designed, built and tested an off-grid, solar powered fridge, with no battery bank. Using an inexpensive chest freezer with a few modifications.

2. [Fridge0: an open source hardware solar fridge with no battery ...](<https://boingboing.net/2019/09/05/frozen-sunshine.html>) Date published: Sep 5, 2019

Joey Hess designed the first Fridge0 a year ago: it uses a standard chest freezer with added thermal mass, a simple controller, ...

...

```
</tool_response>
```

### Step 6

```
<think>
```

To fully verify that Joey Hess matches all the biographical clues in the question, I need to establish whether he has personal memories of a developer's conference in Edinburgh and is also known for enjoying caving. These two details are particularly distinctive and would help confirm his identity as the subject of the question, especially since I already have strong evidence about his rustic lifestyle and solar-powered fridge...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess developer's conference Edinburgh memories", "Joey Hess caving"]}}
```

```
</tool_call>
```

```
<tool_response>
```

```
<tool_call>
```

```
{"name": "ziyaret", "arguments": {"url": ["https://lwn.net/Articles/672352"], "goal": "Yazılım geliştiriciyi, güneş enerjili bir buzdolabı tasarladığı iddiasını, kırsal yaşam tarzı bahsini ve erken bilgisayar alımlarıyla ilgili herhangi bir ayrıntıyı belirle"}}
```

```
</tool_call>
```

```
<tool_response>
```

<https://lwn.net/Articles/672352/> adresindeki, kullanıcı hedefi olan yazılım geliştiriciyi, güneş enerjili bir buzdolabı tasarladığı iddiasını, kırsal yaşam tarzı bahsini ve erken bilgisayar alımlarıyla ilgili herhangi bir ayrıntıyı belirlemeye yönelik faydalı bilgiler şunlardır: ...

```
</tool_response>
```

### Adım 5

```
<think>
```

LWN.net'teki röportaj, Joey Hess'in söz konusu yazılım geliştirici olduğuna dair güçlü kanıtlar sunarak, onun "haritadaki bir delikteki" kırsal yaşam tarzını ve erken bilgisayar deneyimlerini detaylandırmaktadır. Ancak, güneş panelleri ve şebekeden bağımsız yaşamdan bahsedilse de, kendisinin güneş enerjili bir buzdolabı tasarlayıp ürettiğini iddia ettiğine dair henüz açık bir doğrulama bulunmamaktadır – bu, sorunun öznesi olarak kimliğini tam olarak doğrulamak için önemli bir ayrıntıdır...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess solar-powered fridge", "Joey Hess solar powered fridge created", "Joey Hess designed solar-powered fridge", "Joey Hess fridge solar"]}}
```

```
</tool_call>
```

```
<tool_response>
```

'Joey Hess güneş enerjili buzdolabı' araması için yapılan bir Google araması 9 sonuç bulmuştur:

1. [Şebekeden Bağımsız, Güneş Enerjili, Sıfır Bataryalı Buzdolabı](<https://www.notechmagazine.com/2019/09/off-grid-solar-powered-zero-battery-refrigerator.html>)

Yayın tarihi: 1 Eylül 2019

Joey Hess, batarya bankası bulunmayan, şebekeden bağımsız, güneş enerjili bir buzdolabı tasarlamış, inşa etmiş ve test etmiştir. Birkaç modifikasyonla birlikte ucuz bir sandık dondurucu kullanmaktadır.

2. [Buzdolabı0: bataryasız, açık kaynak donanımlı bir güneş enerjili buzdolabı ...](<https://boingboing.net/2019/09/05/frozen-sunshine.html>) Yayınlama tarihi: 5 Eylül 2019

Joey Hess, ilk Buzdolabı0'ı bir yıl önce tasarlamıştır: standart bir sandık dondurucuya ek termal kütle, basit bir kontrolör ve ...

...

```
</tool_response>
```

### Adım 6

```
<think>
```

Joey Hess'in sorudaki tüm biyografik ipuçlarına uyduğunu tam olarak doğrulamak için, Edinburgh'daki bir geliştirici konferansından kişisel anıları olup olmadığını ve aynı zamanda mağaracılıktan hoşlanmasıyla tanınıp tanınmadığını tespit etmem gerekmektedir. Bu iki ayrıntı özellikle ayırt edicidir ve onun sorunun konusu olarak kimliğini doğrulamaya yardımcı olacaktır, zira kırsal yaşam tarzı ve güneş enerjili buzdolabı hakkında zaten güçlü kanıtlara sahibim...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess geliştirici konferansı Edinburgh anılar", "Joey Hess mağaracılık"]}}
```

```
</tool_call>
```

```
<tool_response>
```

A Google search for 'Joey Hess developer's conference Edinburgh memories' found 9 results:  
1. [Leaving Debian With Joey Hess - CoRecursive Podcast](<https://corecursive.com/leaving-debian/>) Date published: Nov 2, 2021

Joey Hess is my guest, and he dedicated significant parts of his adult life to working on Debian. He's going to share what that was like.

2. [Interviews with FLOSS developers: Joey Hess | i.am.z.grim.shell](<https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html>) Date published: Feb 28, 2015

me: Some memorable moments from Debian conferences? joeyh: There are so many! Picnicing on berries and tamales at the Portland farmer's market ...

...

```
</tool_response>
```

#### Step 7

```
<think>
```

To confidently establish whether Joey Hess has personal memories of a developer's conference in Edinburgh—an important detail for confirming he matches all the unique clues in the question—I need direct evidence of his experiences or references to such an event. The interview linked at zgrimshell.github.io specifically asks Joey Hess about memorable moments from Debian conferences...

```
</think>
```

```
<tool_call>
```

```
{"name": "visit", "arguments": {"url": ["https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html"], "goal": "Find mention of Edinburgh and conference memories"}}
```

```
</tool_call>
```

```
<tool_response>
```

The useful information in <https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html> for user goal Find mention of Edinburgh and conference memories as follows:...

```
</tool_response>
```

#### Step 8

```
<think>
```

While I have strong evidence linking Joey Hess to the rustic lifestyle, solar-powered fridge, and developer conference memories, the detail about caving is still unconfirmed. Since "caving" could be a metaphor or refer to an actual hobby, I want to directly verify whether Joey Hess has publicly mentioned caving or spelunking as an interest or activity...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess caving", "Joey Hess cave", "Joey Hess spelunking"]}}
```

```
</tool_call>
```

```
<tool_response>
```

A Google search for 'Joey Hess caving' found 9 results:

1. [Joe Hess on X: ""So come out of your cave walking on your hands ...](<https://twitter.com/RunningJoeInJax/status/104570999537799168>)

"So come out of your cave walking on your hands, and see the world hanging upside down..." #TheCave #Muford&Sons. 3:10 PM · Aug 19, 2011.

2. [Joe Hess (@jhessacx) • Instagram photos and videos](<https://www.instagram.com/jhessacx/>) Day 5 - Kingdom Church and 3 baptisms in the Danube River. The testimonies of the baptism candidates were heartfelt to witness and the after baptism prayers ...

...

'Joey Hess geliştiriciler konferansı Edinburgh anıları' için yapılan bir Google araması 9 sonuç buldu:  
1. [Joey Hess ile Debian'dan Ayrılma - CoRecursive Podcast](<https://corecursive.com/leaving-debian/>) Yayınlama Tarihi: 2 Kasım 2021

Joey Hess, benim misafirimdir ve yetişkin yaşamının önemli bir kısmını Debian üzerinde çalışmaya adamıştır. Bunun nasıl bir deneyim olduğunu paylaşacak.

2. [FLOSS geliştiricileri ile söyleşiler: Joey Hess | i.am.z.grim.shell](<https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html>) Yayınlama Tarihi: 28 Şubat 2015

ben: Debian konferanslarından akılda kalıcı anılar? joeyh: O kadar çok ki! Portland çiftçi pazarında böğürtlen ve tamale pikniği...

...

```
</tool_response>
```

#### Adım 7

```
<think>
```

Joey Hess'in Edinburgh'daki bir geliştiriciler konferansına dair kişisel anıları olup olmadığını kesin olarak belirlemek —sorudaki tüm benzersiz ipuçlarıyla eşleştiğini doğrulamak için önemli bir detaydır— adına, bu tür bir etkinliğe ilişkin deneyimlerine veya atıflara dair doğrudan kanıt ihtiyacım var. [zgrimshell.github.io](https://zgrimshell.github.io) adresinde bağlantısı verilen röportaj, Joey Hess'e özellikle Debian konferanslarından unutulmaz anıları soruyor...

```
</think>
```

```
<tool_call>
```

```
{"name": "ziyaret", "arguments": {"url": ["https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html"], "goal": "Edinburgh ve konferans anılarından bahsedildiğini bul"}}
```

```
</tool_call>
```

```
<tool_response>
```

Kullanıcı hedefi olan Edinburgh ve konferans anılarından bahsedilen bilgiler <https://zgrimshell.github.io/posts/interviews-with-floss-developers-joeys-hess.html> adresinde şu şekildedir:...

```
</tool_response>
```

#### Adım 8

```
<think>
```

Joey Hess'i rustik yaşam tarzına, güneş enerjili buz dolabına ve geliştirici konferansı anılarına bağlayan güçlü kanıtlarım olsa da, mağaracılıkla ilgili detay hala doğrulanmamıştır. "Mağaracılık" bir metafor olabileceği veya gerçek bir hobije atıfta bulunabileceğinin, Joey Hess'in mağaracılığı veya speleolojiye olan ilgisini ya da bu konudaki bir aktivitesini kamuoyunda belirtip belirtmediğini doğrudan doğrulamak istiyorum...

```
</think>
```

```
<tool_call>
```

```
{"name": "search", "arguments": {"query": ["Joey Hess caving", "Joey Hess cave", "Joey Hess spelunking"]}}
```

```
</tool_call>
```

```
<tool_response>
```

'Joey Hess mağaracılık' için yapılan bir Google aramasında 9 sonuç bulundu:

1. [X'teki Joe Hess: ""Yani mağarandan ellerinin üzerinde yürüyerek çıkış...](<https://twitter.com/RunningJoeInJax/status/104570999537799168>)

"Yani mağarandan ellerinin üzerinde yürüyerek çıkış ve dünyayı baş aşağı asılı gör..." #TheCave #Muford&Sons. 15:10 · 19 Ağustos 2011.

2. [Joe Hess (@jhessacx) • Instagram fotoğraf ve videoları](<https://www.instagram.com/jhessacx/>) 5. Gün - Kingdom Kilisesi ve Tuna Nehri'nde 3 vafit. Vafit adaylarının tanıklıklarına şahit olmak yürek ıstıdı ve vafit sonrası dualar...

...

```
</tool_response>
```

### Step 9

```
<think>
```

To determine the make and model of the first computer purchased jointly with his father, I need to locate a source where Joey Hess has directly discussed this formative experience, ideally in his own words. The blog post at the provided URL is likely to contain personal anecdotes and technical details about his early computing history, which is precisely the information required to answer the question...

```
</think>
```

```
<tool_call>
```

```
{"name": "visit", "arguments": {"url": ["http://joeyh.name/blog/entry/atari"]}, "goal": "Find the make and model of the first computer purchased with his father in the 1980s"}
```

```
</tool_call>
```

```
<tool_response>
```

The useful information in <http://joeyh.name/blog/entry/atari/> for user goal Find the make and model of the first computer purchased with his father in the 1980s as follows: ...

```
</tool_response>
```

### Step 10

```
<think>
```

To arrive at the answer "Atari 130XE", I first needed to identify the software developer described in the question by piecing together several unique biographical clues: designing a solar-powered fridge, living a rustic lifestyle "in a hole in the map", having memories from a developer's conference in Edinburgh, and enjoying caving. My searches led me to Joey Hess, whose interviews and blog posts explicitly confirm his off-grid living, the creation of the "fridge0" solar-powered fridge, and fond recollections of Debian conferences—including an event in Edinburgh. To answer the core question about the first computer purchased with his father in the 1980s, I located Joey's blog entry detailing that it was an Atari 130XE, bought jointly with his dad around 1986. This direct evidence from his own writing allowed me to confidently conclude that the Atari 130XE was the correct answer.

```
</think>
```

```
<answer>The first computer they bought together was an Atari 130XE. <answer> ✓
```

```
</tool_response>
```

### Adım 9

```
<think>
```

Babasıyla ortaklaşa satın aldığı ilk bilgisayarın marka ve modelini belirlemek için Joey Hess'in bu şekillendirici deneyimi doğrudan, ideal olarak kendi sözleriyle tartıştığı bir kaynak bulmam gerekiyor. Sağlanan URL'deki blog yazısı, erken bilgisayar geçmişinde kişisel anekdotlar ve teknik detaylar içermesi muhtemel olup, bu da soruyu yanıtlamak için tam olarak gereken bilgiyi sağlamaktadır...

```
</think>
```

```
<tool_call>
```

```
{"name": "ziyaret", "arguments": {"url": ["http://joeyh.name/blog/entry/atari"]}, "goal": "1980'lerde babasıyla birlikte satın alınan ilk bilgisayarın markasını ve modelini bulun"}
```

```
</tool_call>
```

```
<tool_response>
```

<http://joeyh.name/blog/entry/atari/> adresindeki kullanıcının hedefi olan '1980'lerde babasıyla birlikte satın alınan ilk bilgisayarın markasını ve modelini bulun' amacıyla yönelik faydalı bilgiler şunlardır: ...

```
</tool_response>
```

### Adım 10

```
<think>
```

"Atari 130XE" cevabına ulaşmak için öncelikle soruda bahsedilen yazılım geliştiriciyi birkaç benzersiz biyografik ipucunu bir araya getirerek tanımlamam gereki: güneş enerjili bir buz dolabı tasarlaması, "haritadaki bir delikte" kırsal bir yaşam tarzı sürmesi, Edinburgh'daki bir geliştirici konferansından anıları olması ve mağaracılıktan hoşlanması. Yaptığım aramalar beni Joey Hess'e ulaştırdı; kendisinin röportajları ve blog gönderileri, şebeke dışı yaşamını, "fridge0" güneş enerjili buz dolabını geliştirdiğini ve Edinburgh'daki bir etkinlik de dahil olmak üzere Debian konferanslarına dair hoş anılarını açıkça teyit ediyordu. 1980'lerde babasıyla birlikte satın aldığı Atari 130XE olduğunu detaylandıran blog girdisini buldum. Kendi yazılarından elde ettiğim bu doğrudan kanıt, Atari 130XE'nin doğru yanıt olduğu sonucuna güvenle ulaşmamı sağladı.

```
</think>
```

```
<answer> Birlikte satın aldıkları ilk bilgisayar bir Atari 130XE idi. <answer> ✓
```

## References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.
- ByteDance Doubao. Doubao, 2025. URL <http://www.doubao.com/>.
- Google. Measuring compositional generalization, 2020. URL <https://research.google/blog/measuring-compositional-generalization/>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps, 2020. URL <https://arxiv.org/abs/2011.01060>.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- Jina.ai. Jina, 2025. URL <https://jina.ai/>.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- Kyle Jurado, Sydney C Ludvigson, and Serena Ng. Measuring uncertainty. *American Economic Review*, 105(3):1177–1216, 2015.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*, 2024.

## Referanslar

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman ve Dan Mané. Yapay zekâ güvenliğinde somut problemler. *arXiv ön yayını arXiv:1606.06565*, 2016.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan ve Shunyu Yao. Fireact: Dil aracı ince ayarına doğru. *arXiv ön yayını arXiv:2310.05915*, 2023.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou ve Cihang Xie. SFT mi yoksa RL mi? R1 benzeri akıl yürütme yeteneğine sahip büyük görme-dil modellerinin eğitimine yönelik erken bir araştırma. *arXiv ön yayını arXiv:2504.11468*, 2025.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman ve diğerleri. Kod üzerinde eğitilmiş büyük dil modellerinin değerlendirilmesi. *arXiv ön yayını arXiv:2107.03374*, 2021.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine ve Yi Ma. SFT ezberler, RL genelleştirir: Temel modelin eğitim sonrası karşılaşır tırmalı bir çalışması. *arXiv ön yayını arXiv:2501.17161*, 2025.
- ByteDance Doubao. Doubao, 2025. URL <http://www.doubao.com/>.
- Google. Bileşimsel genellemenin ölçülmesi, 2020. URL <https://research.google/blog/measuring-compositional-generalization/>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi ve diğerleri. DeepSeek-R1: LLM'lerde pekiştirmeli öğrenme yoluyla akıl yürütme yeteneğini teşvik etme. *arXiv ön yayını arXiv:2501.12948*, 2025.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara ve Akiko Aizawa. Akıl yürütme adımlarının kapsamlı değerlendirme için çok aşamalı bir QA veri seti oluşturma, 2020. URL <https://arxiv.org/abs/2011.01060>.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang ve Heung-Yeung Shum. Open-reasoner-zero: Temel model üzerinde pekiştirmeli öğrenmeyi öbeklendirmek için açık kaynaklı bir yaklaşım. *arXiv ön yayını arXiv:2503.24290*, 2025.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu ve Lei Ma. Atlamadan önce bak: Büyük dil modelleri için belirsizlik ölçümünün keşifsel bir çalışması. *arXiv ön yayını arXiv:2307.10236*, 2023.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani ve Jiawei Han. Search-r1: Pekiştirmeli öğrenme ile arama motorlarını kullanarak akıl yürütmek ve faydalananmak için LLM'leri eğitmek. *arXiv ön yayını arXiv:2503.09516*, 2025.
- Jina.ai. Jina, 2025. URL <https://jina.ai/>.
- Mandar Joshi, Eunsol Choi, Daniel S Weld ve Luke Zettlemoyer. Triviaqa: Okuduğunu anlaması için geniş ölçekli, uzaktan denetimli bir meydan okuma veri kümlesi. *arXiv ön yayını arXiv:1705.03551*, 2017.
- Kyle Jurado, Sydney C Ludvigson ve Serena Ng. Belirsizliği ölçmek. *American Economic Review*, 105(3):1177–1216, 2015.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum ve Andrew Gordon Wilson. Büyük dil modellerine bilmediklerini bilmeleri öğretilmelidir. *arXiv ön yayını arXiv:2406.08391*, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. Lara: Benchmarking retrieval-augmented generation and long-context llms—no silver bullet for lc or rag routing. *arXiv preprint arXiv:2502.09977*, 2025a.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025b.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776, 2025c. doi: 10.48550/ARXIV.2504.21776. URL <https://doi.org/10.48550/arXiv.2504.21776>.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Calibrating llm-based evaluator. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 2638–2656. ELRA and ICCL, 2024. URL <https://aclanthology.org/2024.lrec-main.237>.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

OpenAI. Hello GPT-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.

OpenAI. Deep research system card, 2025a. URL <https://cdn.openai.com/deep-research-system-card.pdf>.

OpenAI. Introducing openai gpt-4.1, 2025b. URL <https://openai.com/index/gpt-4-1/>.

OpenAI. Introducing openai o3 and o4-mini, 2025c. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.

OpenAI. Introducing simpleqa, 2025d. URL <https://openai.com/index/introducing-simpleqa/>.

Qwen Team. QwQ-32B: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Doğal sorular: soru yanıtlama araştırmaları için bir çalışma ölçütü. *Hesaplamalı Dilbilim Derneği İşlemleri*, 7:453–466, 2019.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang ve Minhao Cheng. Lara: Geri çağırma-artırılmış üretim ve uzun bağlılı BDM'lerin karşılaştırması – LC veya RAG yönlendirmesi için tek çözüm yok. *arXiv ön yayını arXiv:2502.09977*, 2025a.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang ve Zhicheng Dou. Search-o1: Ajan destekli arama ile geliştirilmiş büyük akıl yürütme modelleri. *arXiv ön yayını arXiv:2501.05366*, 2025b.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen ve Zhicheng Dou. Webthinker: Büyük Akıl Yürütme Modellerini Derin Araştırma Yeteneğiyle Güçlendirmek. *CoRR*, abs/2504.21776, 2025c. doi: 10.48550/ARXIV.2504.21776. URL <https://doi.org/10.48550/arXiv.2504.21776>.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun ve Qi Zhang. LLM Tabanlı Değerleyicinin Kalibrasyonu. Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti ve Nianwen Xue (eds.), *2024 Ortak Uluslararası Hesaplamalı Dilbilim, Dil Kaynakları ve Değerlendirme Konferansı Bildirileri, LREC/COLING 2024, 20-25 Mayıs, 2024, Torino, İtalya*, pp. 2638–2656. ELRA ve ICCL, 2024. URL <https://aclanthology.org/2024.lrec-main.237>.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun ve Thomas Scialom. Gaia: genel yapay zeka asistanları için bir kıyaslama. İçinde *On İlk Uluslararası Öğrenme Temsilleri Konferansı*, 2023.

OpenAI. Merhaba GPT-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.

OpenAI. Derin araştırma sistem kartı, 2025a. URL <https://cdn.openai.com/deep-research-system-card.pdf>.

OpenAI. OpenAI GPT-4.1 Tanıtımı, 2025b. URL <https://openai.com/index/gpt-4-1/>.

OpenAI. OpenAI O3 ve O4-mini Tanıtımı, 2025c. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.

OpenAI. simpleqa tanıtılıyor, 2025d. URL <https://openai.com/index/introducing-simpleqa/>.

Qwen Ekibi. QwQ-32B: Pekiştirmeli öğrenmenin gücünü kucaklamak, Mart 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Açık dil modellerinde matematiksel akıl yürütmenin sınırlarını zorlamak. *arXiv ön yayını arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin ve Chuan Wu. Hybridflow: Esnek ve verimli bir rlhf çerçevesi. Yirminci Avrupa Bilgisayar Sistemleri Konferansı Bildirileri, ss. 1279–1297, 2025.

Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper ve Bryan Catanzaro. Megatron-lm: Model paralelliğini kullanarak çok milyar parametreli dil modellerinin eğitimi. *arXiv ön yayını arXiv:1909.08053*, 2019.

- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.
- Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.
- Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri, and Dawn Song. Climbing the ladder of reasoning: What llms can-and still can't-solve after sft? *arXiv preprint arXiv:2504.11741*, 2025.
- Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*, 2025.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.
- Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. Leave no document behind: Benchmarking long-context llms with extended multi-doc QA. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5627–5646. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.emnlp-main.322>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecamp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.
- Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36:6941–6960, 2023.
- Tom D Wilson. Models in information behaviour research. *Journal of documentation*, 55(3):249–270, 1999.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025a.
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal, 2025b. URL <https://arxiv.org/abs/2501.07572>.
- x.ai. Grok 3 beta — the age of reasoning agents, 2025. URL <https://x.ai/news/grok-3>.
- Xbench-Team. Xbench-deepsearch, 2025. URL <https://xbench.org/agi/aisearch>.
- Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang ve Ji -Rong Wen. R1-searcher: Pekiştirmeli öğrenme yoluyla büyük dil modellerinde arama yeteneğini teşvik etmek. *arXiv ön yayını arXiv:2503.05592*, 2025.
- Chuanneng Sun, Songjun Huang ve Dario Pompili. Büyük dil modeli tabanlı çoklu ajan pekiştirmeli öğrenme: Mevcut ve gelecekteki yönelikler. *arXiv ön yayını arXiv:2405.11106*, 2024.
- Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri ve Dawn Song. Akıl yürütme merdivenini tırnamak: LLM'ler SFT'den sonra neler çözülebilir ve hala neleri çözmez? *arXiv ön yayını arXiv:2504.11741*, 2025.
- Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu ve J Andrew Bagnell. Tüm yollar ola- sılığa çıkar: İnce ayarda pekiştirmeli öğrenmenin değeri. *arXiv ön yayını arXiv:2503.01067*, 2025.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot ve Ashish Sabharwal. Musique: Tek atlama- malı soru birleşimiyle çok atlamalı sorular. *Hesaplamalı Dilbilim Derneği İşlemleri*, 10:539–554, 2022.
- Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang ve Yongbin Li. Hiçbir belgeyi geride bırakmayın: Uzatılmış çok belgeli QA ile uzun bağlımlı BDM'leri kıyaslama. Yaser Al-Onaizan, Mohit Bansal ve Yun-Nung Chen (eds.), *2024 Doğal Dil İşlemede Empirik Yöntemler Konferansı Bil- dirileri, EMNLP 2024, Miami, FL, ABD, 12-16 Kasım 2024*, ss. 5627–5646. Hesaplamalı Dilbilim Derneği, 2024. URL <https://aclanthology.org/2024.emnlp-main.322>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, vd. Düşünce zinciri istemi, büyük dil modellerinde akıl yürütme tetikler. *Sınırsız bilgi işleme sistemlerindeki gelişmeler*, 35:24824–24837, 2022.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman ve William Fedus. Büyük dil modellerinde kısa biçimli gerçekliğin ölçülmesi. *arXiv ön yayını arXiv:2411.04368*, 2024.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus ve Amelia Glaese. Browsecamp: Tarama ajanları için basit ama zorlu bir kıyaslama. *arXiv ön yayını arXiv:2504.12516*, 2025.
- Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge ve Wieland Brendel. İlk prensip- lerden bileşimsel genelleme. *Sınırsız Bilgi İşleme Sistemlerindeki Gelişmeler*, 36:6941–6960, 2023.
- Tom D Wilson. Bilgi davranış araştırmalarında modeller. *Dokümantasyon Dergisi*, 55(3):249–270, 1999.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, et al. Webdancer: Otonom bilgi arama ajanlığına doğru. *arXiv ön yayını arXiv:2505.22648*, 2025a.
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie ve Fei Huang. Webgeçişinde LLM'leri kıyaslama, 2025b. URL <https://arxiv.org/abs/2501.07572>.
- x.ai. Grok 3 beta — akıl yürütme ajanları çağrı, 2025. URL <https://x.ai/news/grok-3>.
- Xbench-Takımı. Xbench-derin arama, 2025. URL <https://xbench.org/agi/aisearch>.
- Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao ve Zheng Ma. Frekans ilkesi: Fourier analizi derin sınırlarını aydınlatıyor. *arXiv ön yayını arXiv:1901.06523*, 2019.

- Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, pp. 1–38, 2024.
- Zhi-Qin John Xu, Yaoyu Zhang, and Zhangchen Zhou. An overview of condensation phenomenon in deep learning. *arXiv preprint arXiv:2504.09484*, 2025.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
- Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Hao Wang, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang, et al. Towards widening the distillation bottleneck for reasoning models. *arXiv e-prints*, pp. arXiv–2503, 2025.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.
- Dingchu Zhang, Yida Zhao, Jialong Wu, Baixuan Li, Wenbiao Yin, Liwen Zhang, Yong Jiang, Yufeng Li, Kewei Tu, Pengjun Xie, and Fei Huang. Evolvesearch: An iterative self-evolving search agent, 2025a. URL <https://arxiv.org/abs/2505.22501>.
- Shaokun Zhang, Yi Dong, Jieyu Zhang, Jan Kautz, Bryan Catanzaro, Andrew Tao, Qingyun Wu, Zhiding Yu, and Guilin Liu. Nemotron-research-tool-n1: Tool-using language models with reinforced reasoning. *arXiv preprint arXiv:2505.00024*, 2025b.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments, 2025. URL <https://arxiv.org/abs/2504.03160>.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, et al. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*, 2025.
- Zhi-Qin John Xu, Yaoyu Zhang ve Tao Luo. Derin öğrenmede frekans ilkesi/spektral yanılık genel bakış. *Uygulamalı Matematik ve Hesaplama Üzerine İletişimler*, ss. 1–38, 2024.
- Zhi-Qin John Xu, Yaoyu Zhang ve Zhangchen Zhou. Derin öğrenmede yoğunlaşma fenomenine genel bakış. *arXiv ön yayını arXiv:2504.09484*, 2025.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei ve diğerleri. Qwen2.5 teknik raporu. *arXiv ön yayını arXiv:2412.15115*, 2024.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov ve Christopher D Manning. Hotpotqa: Çeşitli, açıklanabilir çok aşamalı soru yanıtlama için bir veri kümesi. *arXiv ön yayını arXiv:1809.09600*, 2018.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan ve Yuan Cao. React: Dil modellerinde muhakeme ve eylemi sinerjik hale getirme. In *Uluslararası Öğrenme Temsilieri Konferansı (ICLR)*, 2023.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia ve Pengfei Liu. Limo: Akıl yürütme için az daha çoktur. *arXiv ön yayını arXiv:2502.03387*, 2025.
- Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Hao Wang, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang ve diğerleri. Akıl yürütme modelleri için damıtma darboğazını genişletmeye yönelik. *arXiv e-printler*, ss. arXiv–2503, 2025.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu ve diğerleri. Dapo: Ölçekli açık kaynaklı bir LLM pekiştirmeli öğrenme sistemi. *arXiv ön yayını arXiv:2503.14476*, 2025.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song ve Gao Huang. Pekiştirmeli öğrenme, temel modelin ötesinde LLM'lerde akıl yürütme kapasitesini gerçekten teşvik ediyor mu? *arXiv ön yayını arXiv:2504.13837*, 2025.
- Dingchu Zhang, Yida Zhao, Jialong Wu, Baixuan Li, Wenbiao Yin, Liwen Zhang, Yong Jiang, Yufeng Li, Kewei Tu, Pengjun Xie ve Fei Huang. Evolvesearch: Tekrarlayan, kendi kendine evrilen bir arama ajanı, 2025a. URL <https://arxiv.org/abs/2505.22501>.
- Shaokun Zhang, Yi Dong, Jieyu Zhang, Jan Kautz, Bryan Catanzaro, Andrew Tao, Qingyun Wu, Zhiding Yu ve Guolin Liu. Nemotron-research-tool-n1: Pekiştirilmiş akıl yürütmeye sahip araç kullanan dil modelleri. *arXiv ön yayını arXiv:2505.00024*, 2025b.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu ve Pengfei Liu. Deepresearcher: Gerçek Dünya Ortamlarında Pekiştirmeli Öğrenme Yoluyla Derin Araştırmaların Ölçeklendirilmesi, 2025. URL <https://arxiv.org/abs/2504.03160>.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao ve diğerleri. Browsecomp-zh: Büyük dil modellerinin Çince web tarama yeteneğinin karşılaştırmalı değerlendirmesi. *arXiv ön yayını arXiv:2504.19314*, 2025.