

MCP vs RAG vs NLWeb vs HTML

A Comparison of the Effectiveness and Efficiency of Different Agent Interfaces to the Web



Aaron Steiner, Ralph Peeters and Christian Bizer

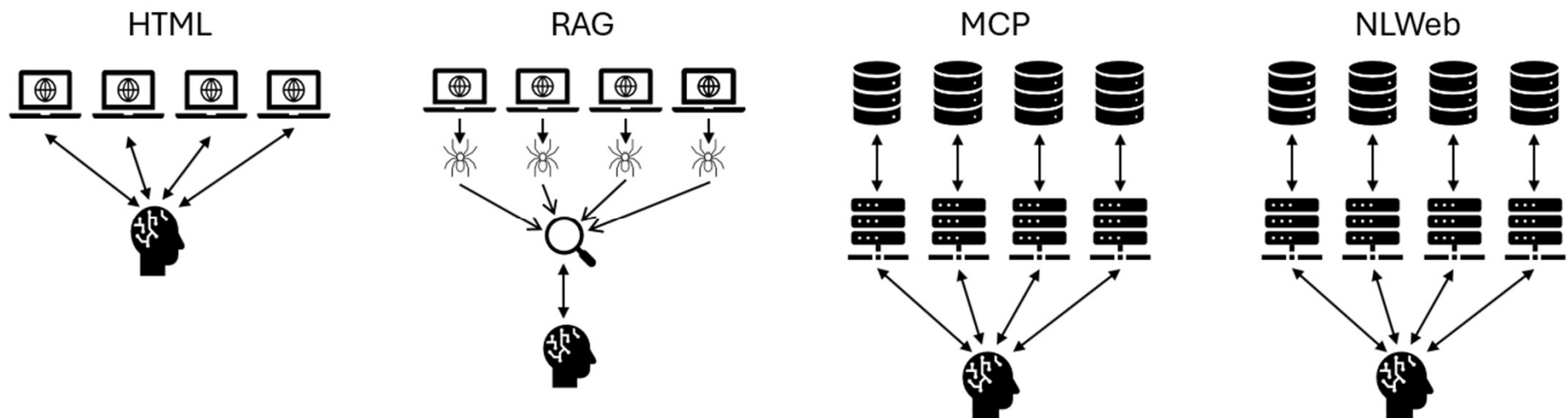
Interfaces to the Web for LLM Agents

- Many different agent tools exist: OpenAI Operator, Manus, anthropic computer use.
- **Agents use different interfaces to interact with websites:**
 - HTML Browsing Agents
 - Web RAG
 - Structured APIs (MCP)
 - Standardized NL APIs (NLWeb)
- **Research question:** To what extent does the interface impact the performance and efficiency of web agents?
- **What we do:** Benchmark LLM agents that use different interfaces on the **same set of e-commerce tasks** (search products, compare products, add to cart, checkout).

Agenda

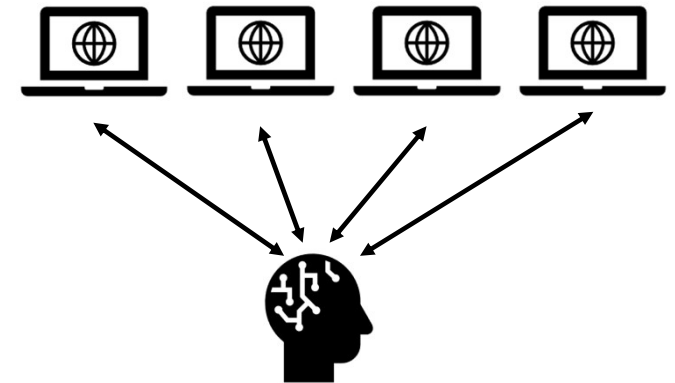
1. Introduction of agent interfaces
2. The WebMall benchmark
3. Experimental setup
4. Results - effectiveness
5. Results - efficiency

Overview: Architectures and Interfaces



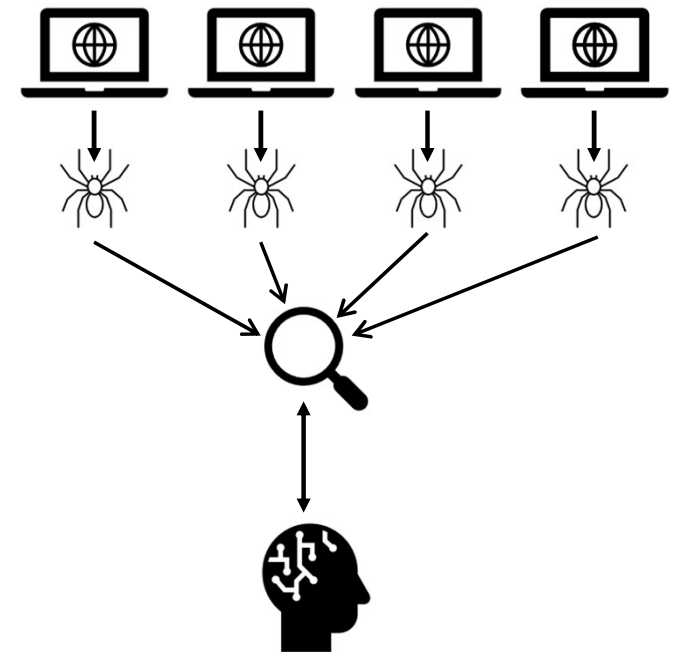
HTML Browsing

- Agents interact with traditional human-oriented HTML pages
- Actions: follow links, click, fill out forms, navigate page-by-page
- Supports interaction with *any* website without requiring APIs
- Strengths: No shop-side changes required
- Limitations:
 - Navigation incurs high overhead
 - Difficulty parsing the observation space



Retrieval-Augmented Generation (RAG)

- Pre-crawls websites and indexes the preprocessed content
- Agent issues search queries → retrieves relevant documents/chunks
- Supports cross-shop retrieval using single index
- Limitations:
 - Snapshot lag
 - No native transaction support



Example: Query and Response

Agent query:

```
{"tool_name": "search_products", "tool_args": {"query": "Canon EOS R5 Mark II", "results": 15}}
```

Tool response:

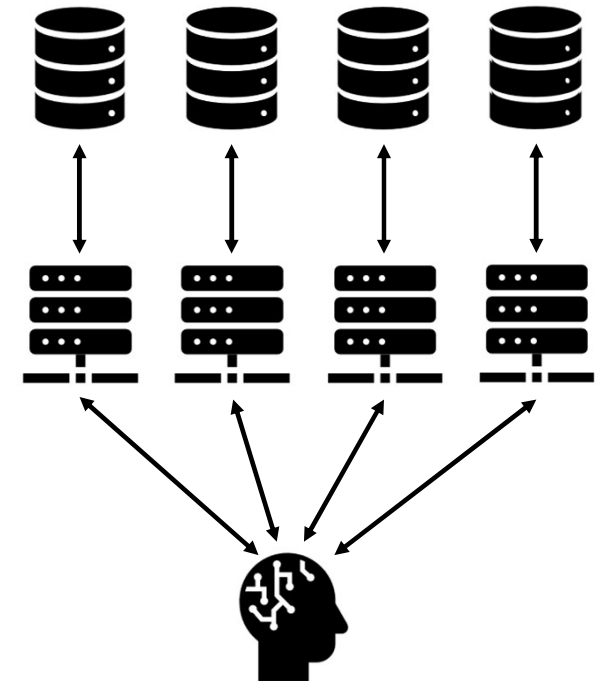
```
"title": "Canon EOS R5 Mark II Mirrorless Digital Camera Body Only",
```

```
"url": "https://webmall-4.de/product/canon-eos-r5-mark-ii-mirrorless-digital-camera-body-only/",
```

```
"description": "Free Shipping on Orders Over $50! Limited Time Offer u2013 Shop Now!nHardware CafennYour cart is currently empty!nSupport:n+1 (129) 366-8351nhardware@cafe.comnCanon EOS R5 Mark II Mirrorless Digital Camera, Body OnlynCanon EOS R5 Mark II Mirrorless Digital Camera, Body Onlyn1.399,00 u20acnCategory: EOS R5 Mark IIInDescriptionnDescriptionnThe EOS R5 Mark II is a worthy successor to the incredibly versatile EOS R5 bringing a host of new features that make this the most powerful EOS 5 series camera ever. It starts with a brand-new 45MP sensor that is back illuminated and stacked for blazingly fast read-out speeds that allow for a 30 fps electronic shutter, then builds on that with major focus improvements, increased video capabilities, and so much more. nDIGIC AcceleratorThe DIGIC Accelerator chip is a completely new front-end processor developed by Canon to supplement the power of the proven DIGIC X processor.[....]"
```


Model Context Protocol (MCP)

- Standardizes how LLMs discover and call external tools (APIs)
- Each Web shop server exposes a set of functions (search, cart, checkout, etc.) → heterogeneity in schemas & responses
- Communication via JSON-RPC → LLM sends tool invocation, server returns structured JSON
- Enables precise actions (e.g., “add_to_cart(product_id=123)”)
- Strengths: structured control, real-time actions
- Limitations:
 - Non-standard endpoints across sites → agent needs to deal with heterogeneity
 - Shops need to implement public API



Example: Query and Response

Agent query:

```
"E-Store Athletes": {"tool_arguments": {"query": "Canon EOS R5 Mark II", "per_page": 20, "page": 1, "include_descriptions": true}}
```

Tool response:

```
{ "ID": "1963", "label": "EOS R5 Mark II Body",  
  "desc": {  
    "longVersion": "<p>The EOS R5 Mark II is a worthy successor  
to the incredibly [...]</p>\n",  
    "quickPitch": "<p>The EOS R5 Mark II [...]</p>\n",  
    "stock": { "itemCode": "1963", "status": "In stock", "leftOverCount": 0},  
    "labels": { "categories": ["Canon"]},  
    "priceInfo": { "current": "1500.0", "usual": ""},  
    "addresses": {  
      "selfLink": "https://webmall-1.de/mark-ii-body/",  
      "shareLink": "https://webmall-1.de/r5-mark-ii-body/"  
    }  
  }
```

Agent query:

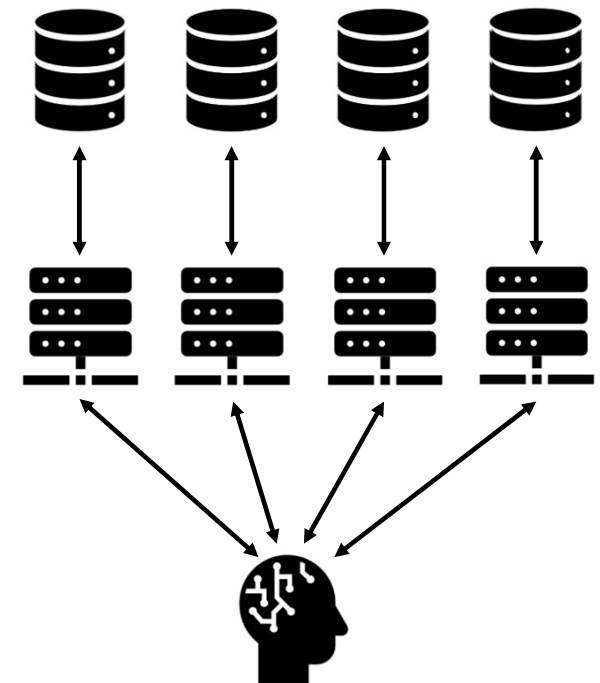
```
"TechTalk": {"tool_arguments": {"query": "Canon EOS R5 Mark II",  
  "limit": 20, "page_num": 1, "sort_by_price": "none",  
  "include_descriptions": true }}
```

Tool response:

```
{ "catalog_entry_id": "3526", "merchandise_title": "Canon EOS R5 II",  
  "financial_details": { "cost_amount": "1544.54", "standard_rate": ""},  
  "content_sections": {  
    "detailed_info": "<p>Full-Frame Stacked BSI CMOS [...]</p>\n",  
    "visual_assets": [ "https://webmall-2.de/wp-content/uploads/0-scaled.jpg"],  
  },  
  "inventory_tracking": { "product_identifier": "3526",  
    "availability_state": "On the shelf"},  
  "classification_tags": [ "Cameras"],  
  "direct_link": "https://webmall-2.informatik.uni-mannheim.de/product/canon-eos-r5-ii/"  
}
```

Natural Language Web (NLWeb)

- Extends the MCP protocol with a uniform, natural-language interface
- Each site hosts an “/ask” endpoint that accepts NL queries (e.g., “Find laptops under €1000 with 16 GB RAM”)
- Responses returned as Schema.org-aligned JSON
 - consistent structure across all providers
- Allows cross-shop aggregation and semantic comparison
- Uses MCP for complementary actions (cart, checkout)
- Strengths: standardized outputs, less heterogeneity
- Limitations: requires site adoption, still experimental



Example: Query and Response

Agent query:

```
"mcp_server": "E-Store Athletes", "tool_arguments": { "question": "Canon EOS R5 Mark II", "top_k": 10 },
```

Tool response:

```
{ "@context": "https://schema.org/", "@type": "Product", "name": "EOS R5 Mark II Body",  
  "image": [ https://webmall-1.de/wp-content/uploads/2025/05/3119.jpg ],  
  "description": "<p>The EOS R5 Mark [...].</p>",  
  "category": "Canon",  
  "offers": {  
    "@type": "Offer",  
    "url": "https://webmall-1.de/product/eos-r5-mark-ii-body/",  
    "priceCurrency": "EUR", "price": "1500.0",  
    "availability": "https://schema.org/InStock",  
    "itemCondition": https://schema.org/NewCondition  
  },  
  "url": "https://webmall-1.de/product/eos-r5-mark-ii-body/",  
  "identifier": { "@type": "PropertyValue", "name": "WooCommerce ID", "value": "1963"}  
}
```

Steiner - MCP vs RAG vs NLWeb vs HTML - 02.12.2025

Agent query:

```
"mcp_server": "TechTalk", "tool_arguments": { "question": "Canon EOS R5 Mark II", "top_k": 10},
```

Tool response:

```
{ "@context": "https://schema.org/", "@type": "Product", "name": "Canon EOS R5 II",  
  "image": ["https://webmall-2.de/wp-content/uploads/2025/05/3120.jpg"],  
  "description": "<p>Pixel Intelligent AF, 60 Raw/4K 60 SRAW/4K 120 10-Bit [...]",  
  "category": "Cameras",  
  "offers": {  
    "@type": "Offer",  
    "url": "https://webmall-2.de/product/canon-eos-r5-ii/",  
    "priceCurrency": "EUR", "price": "1544.54",  
    "availability": "https://schema.org/InStock",  
    "itemCondition": https://schema.org/NewCondition  
  },  
  "url": "https://webmall-2.de/product/canon-eos-r5-ii/",  
  "identifier": { "@type": "PropertyValue", "name": "WooCommerce ID", "value": "3526"}  
}
```

11

Interface comparison

Aspect	HTML	RAG	MCP	NLWeb
E-Shops				
Interface	HTML pages	Retrial API	Proprietary APIs	Standardized API
Search functionality	Free-text search per shop	Search engine, crawled content	Per shop index; structured data	Per shop index; structured data
Search response	HTML result list including links	Pre-processed HTML pages	Heterogeneous JSON	Schema.org JSON
Agent				
Communication protocol	HTML over HTTP	Direct calls to search engine	JSON-RPC via MCP	JSON-RPC via MCP
Query strategy	Site search and browsing	Multi-query	Multi-query per shop	Multi-query per shop
Query refinement	Interactive page exploration	Self-evaluation & iteration	Self-evaluation & iteration	Self-evaluation & iteration
Add to cart / checkout	Clicking & form filling	Direct function calls	MCP tool invocation	MCP tool invocation

Agenda

- Introduction of agent interfaces
- **WebMall Benchmark**
- Experimental setup
- Results - Effectiveness
- Results - Efficiency

WebMall: Online Shopping Benchmark

- **Environment:** Simulates an online shopping environment consisting of 4 online stores and associated product data
 - With **different layouts**
 - Containing **heterogeneous product offers**
- **Task Set:** Contains definitions for tasks in **4 task categories** consisting of
 - **Instructions** for the agent
 - **Expected answers** (definition of success)

Specific & Vague Product Search

Specific Product Search (23 tasks)

- User intent is fully specified (exact model or clear attributes).
- Challenge: correctly interpret constraints and retrieve all matching offers.

Example:

Find all offers for the AMD Ryzen 9 5900X.

Vague Product Search (19 tasks)

- Requirements are open-ended or only partially defined.
- Challenge: explore broadly, interpret potential meanings, and iteratively refine the search

Example:

Find all offers for compatible CPUs for this motherboard: {PRODUCT_URL}.

Cheapest Product Search & Transactional Tasks

Cheapest Product Search (26 tasks)

- Combines retrieval with a global price constraint.
- Challenge: identify all valid products, then select the lowest-priced correct offer.

Example:

Find the cheapest offers for each model of mid-tier Nvidia gaming GPUs in the 4000 series.

Action & Transaction (15 tasks)

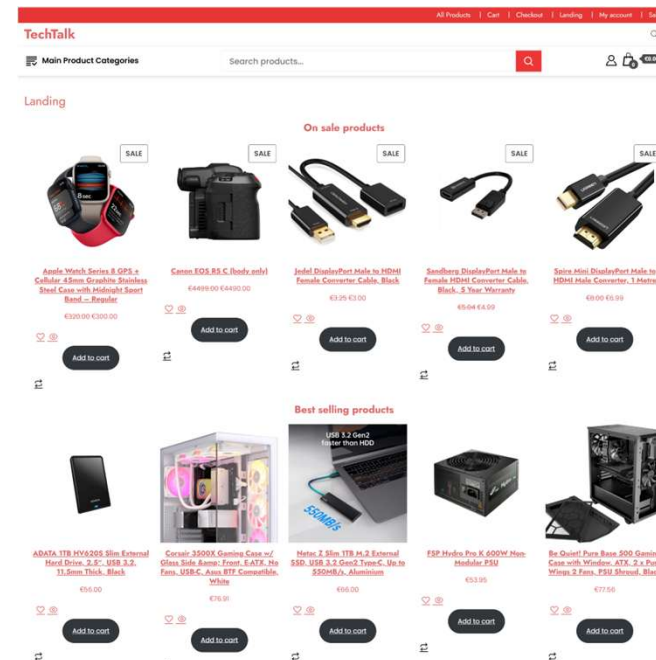
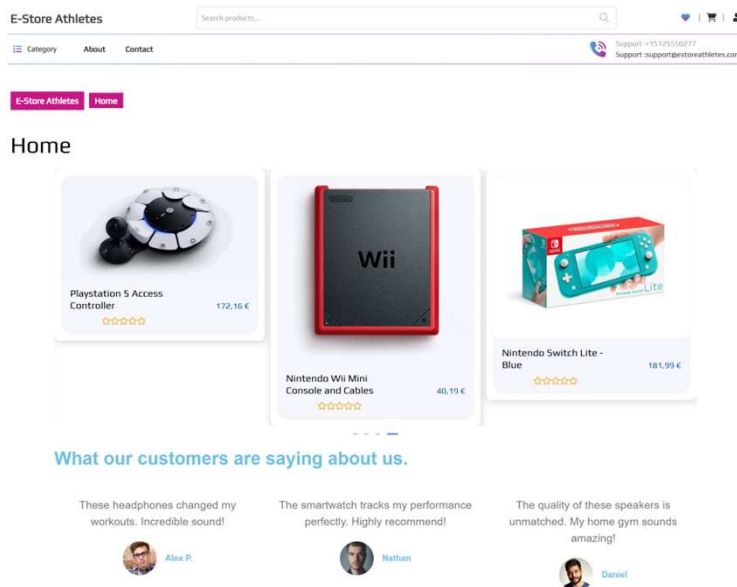
- Add-to-cart and checkout operations across one or multiple shops.
- Challenge: execute multi-step workflows and coordinate actions correctly.

Example:

Add the product on page {url} to the shopping cart and complete the checkout process.

WebMall: The Environment

4 shops created with the WordPress plugin WooCommerce



Experimental Setup

- Evaluated model: GPT-4.1, GPT-5, GPT-5-mini, Claude Sonnet 4
- CR, F1 as evaluation metric for performance
 - Completion rate: Measures exact match between agent answer and test set.
 - F1: Measures the overlap between the products returned by an agent and the ground-truth set. → Partial completion.
 - All reported results are averaged across task groups and models. Tables show these aggregated averages, and the best-performing agent-model combination is highlighted.
- Efficiency measured in token usage and cost

Effectiveness

Agent	CR	F1	Token	Cost	Runtime
HTML	0.57	0.67	241,136	\$0.52	291 s
RAG	0.68	0.77	46,667	\$0.10	50 s
MCP	0.62	0.75	139,569	<u>\$0.27</u>	62 s
NLWeb	<u>0.64</u>	<u>0.76</u>	<u>71,214</u>	\$0.10	<u>53 s</u>

* Results are averaged across all tasks and models and aggregated by interface.

- **RAG** attains top performance
- **NLWeb** close, **MCP** comparable
- **HTML** trails by **≥ 9 percentage points ($\approx +0.09$ F1)**

Model Effectiveness

Model	F1
GPT-4.1	0.70
GPT-5	0.82
GPT-5-mini	<u>0.72</u>
Claude 4 Sonnet	<u>0.72</u>

- Model choice has considerable performance implications
- GPT 5 attains best overall performance

→ **Highest overall performance: RAG with GPT5: 87 F1**

* Results are averaged across all tasks and interfaces and aggregated by model.

A Closer Examination of Effectiveness

- **Specific Product Search (23 tasks)**
 - Find all offers for the AMD Ryzen 9 5900X.
 - Find all offers for Fractal Design PC Gaming Cases which support 240mm radiators and 330mm GPUs.
- **Vague Product Search (19 tasks)**
 - Find an offer for a large PC case from ASUS that can be easily carried to frequent LAN parties.
 - Find all offers for compatible CPUs for this motherboard: {PRODUCT_URL}.
- **Cheapest Product Search (26 tasks)**
 - Find the cheapest offer for a Be Quiet! Pure Base 500 Gaming Case.
 - Find the cheapest offers for each model of mid-tier nVidia gaming GPUs in the 4000 series.
- **Action & Transaction (15 tasks)**
 - Find all offers for the Asus DUAL RTX4070 SUPER OC White and add each of them to the respective shopping cart.
 - Add the product on page {url} to the shopping cart and complete the checkout process.

Specific Product Search

- RAG and NLWeb achieve the best overall performance
- API MCP showcases comparable performance
- HTML trails by more than 10 F1

→ Highest overall performance RAG, API MCP and NLWeb with GPT-5: 96% F1

Interface	F1
HTML	0.77
RAG	0.92
API MCP	<u>0.90</u>
NLWeb	0.92

* Results are averaged across models and aggregated by interface.

A Closer Examination of Effectiveness

- **Specific Product Search (23 tasks)**
 - Find all offers for the AMD Ryzen 9 5900X.
 - Find all offers for Fractal Design PC Gaming Cases which support 240mm radiators and 330mm GPUs.
- **Vague Product Search (19 tasks)**
 - Find an offer for a large PC case from ASUS that can be easily carried to frequent LAN parties.
 - Find all offers for compatible CPUs for this motherboard: {PRODUCT_URL}.
- **Cheapest Product Search (26 tasks)**
 - Find the cheapest offer for a Be Quiet! Pure Base 500 Gaming Case.
 - Find the cheapest offers for each model of mid-tier nVidia gaming GPUs in the 4000 series.
- **Action & Transaction (15 tasks)**
 - Find all offers for the Asus DUAL RTX4070 SUPER OC White and add each of them to the respective shopping cart.
 - Add the product on page {url} to the shopping cart and complete the checkout process.

Vague Product Search

- NLWeb performs best
- HTML and RAG showcase similar performance

Interface	F1
HTML	<u>0.63</u>
RAG	<u>0.63</u>
API MCP	0.59
NLWeb	0.66

* Results are averaged across models and aggregated by interface.

Results for Vague Product Search

Agent	Model	CR	F1	Token	Cost	Runtime
HTML	GPT-4.1	0.32	0.49	157,639	\$0.34	97 s
HTML	GPT-5	0.60	<u>0.78</u>	288,320	\$0.60	641 s
HTML	GPT-5-mini	0.32	0.60	255,852	\$0.09	248 s
HTML	Sonnet 4	0.42	0.60	337,967	\$1.25	361 s
RAG	GPT-4.1	0.26	0.53	18,198	\$0.04	9 s
RAG	GPT-5	0.74	0.82	111,872	\$0.20	147 s
RAG	GPT-5-mini	0.58	0.66	68,452	\$0.02	83 s
RAG	Sonnet 4	0.37	0.41	92,409	\$0.30	42 s
MCP	GPT-4.1	0.11	0.46	27,456	\$0.06	<u>12 s</u>
MCP	GPT-5	0.47	0.65	154,716	\$0.23	119 s
MCP	GPT-5-mini	0.53	0.73	105,082	<u>\$0.03</u>	101 s
MCP	Sonnet 4	0.32	0.53	276,672	\$0.85	53 s
NLWeb	GPT-4.1	0.37	0.60	<u>27,689</u>	\$0.06	13 s
NLWeb	GPT-5	0.53	0.72	77,641	\$0.12	62 s
NLWeb	GPT-5-mini	<u>0.63</u>	0.77	119,435	\$0.04	126 s
NLWeb	Sonnet 4	0.37	0.55	62,435	\$0.20	26 s

→ Reasoning helps significantly in this task set with GPT-5 attaining a 82% F1 using the HTML and RAG interfaces

A Closer Examination of Effectiveness

- **Specific Product Search (23 tasks)**
 - Find all offers for the AMD Ryzen 9 5900X.
 - Find all offers for Fractal Design PC Gaming Cases which support 240mm radiators and 330mm GPUs.
- **Vague Product Search (19 tasks)**
 - Find an offer for a large PC case from ASUS that can be easily carried to frequent LAN parties.
 - Find all offers for compatible CPUs for this motherboard: {PRODUCT_URL}.
- **Cheapest Product Search (26 tasks)**
 - **Find the cheapest offer for a Be Quiet! Pure Base 500 Gaming Case.**
 - **Find the cheapest offers for each model of mid-tier nVidia gaming GPUs in the 4000 series.**
- **Action & Transaction (15 tasks)**
 - Find all offers for the Asus DUAL RTX4070 SUPER OC White and add each of them to the respective shopping cart.
 - Add the product on page {url} to the shopping cart and complete the checkout process.

Cheapest Product Search

- HTML, RAG and API MCP all showcase similar drops in performance
- NLWeb shows larger performance impacts

→ Price constraints effect performance on average by 11 % F1. NLWeb most effected 19 % reduction.

→ Highest score GPT-5 using RAG 78% F1

Interface	F1
HTML	0.61
RAG	0.68
MCP	<u>0.63</u>
NLWeb	0.60

* Results are averaged across models and aggregated by interface.

A Closer Examination of Effectiveness

- **Specific Product Search (23 tasks)**
 - Find all offers for the AMD Ryzen 9 5900X.
 - Find all offers for Fractal Design PC Gaming Cases which support 240mm radiators and 330mm GPUs.
- **Vague Product Search (19 tasks)**
 - Find an offer for a large PC case from ASUS that can be easily carried to frequent LAN parties.
 - Find all offers for compatible CPUs for this motherboard: {PRODUCT_URL}.
- **Cheapest Product Search (26 tasks)**
 - Find the cheapest offer for a Be Quiet! Pure Base 500 Gaming Case.
 - Find the cheapest offers for each model of mid-tier nVidia gaming GPUs in the 4000 series.
- **Action & Transaction (15 tasks)**
 - Find all offers for the Asus DUAL RTX4070 SUPER OC White and add each of them to the respective shopping cart.
 - Add the product on page {url} to the shopping cart and complete the checkout process.

Action & Transaction

- API MCP and NLWeb attain best performance
 - HTML performance is negatively affected by GPT-5 models
 - Automation of workflows via agents viable
- Overall best performance achieved by GPT-4.1 using HTML 1.00 F1

Interface	F1
HTML	0.77
RAG	0.86
MCP	<u>0.93</u>
NLWeb	0.95

* Results are averaged across models and aggregated by interface.

Efficiency and token usage

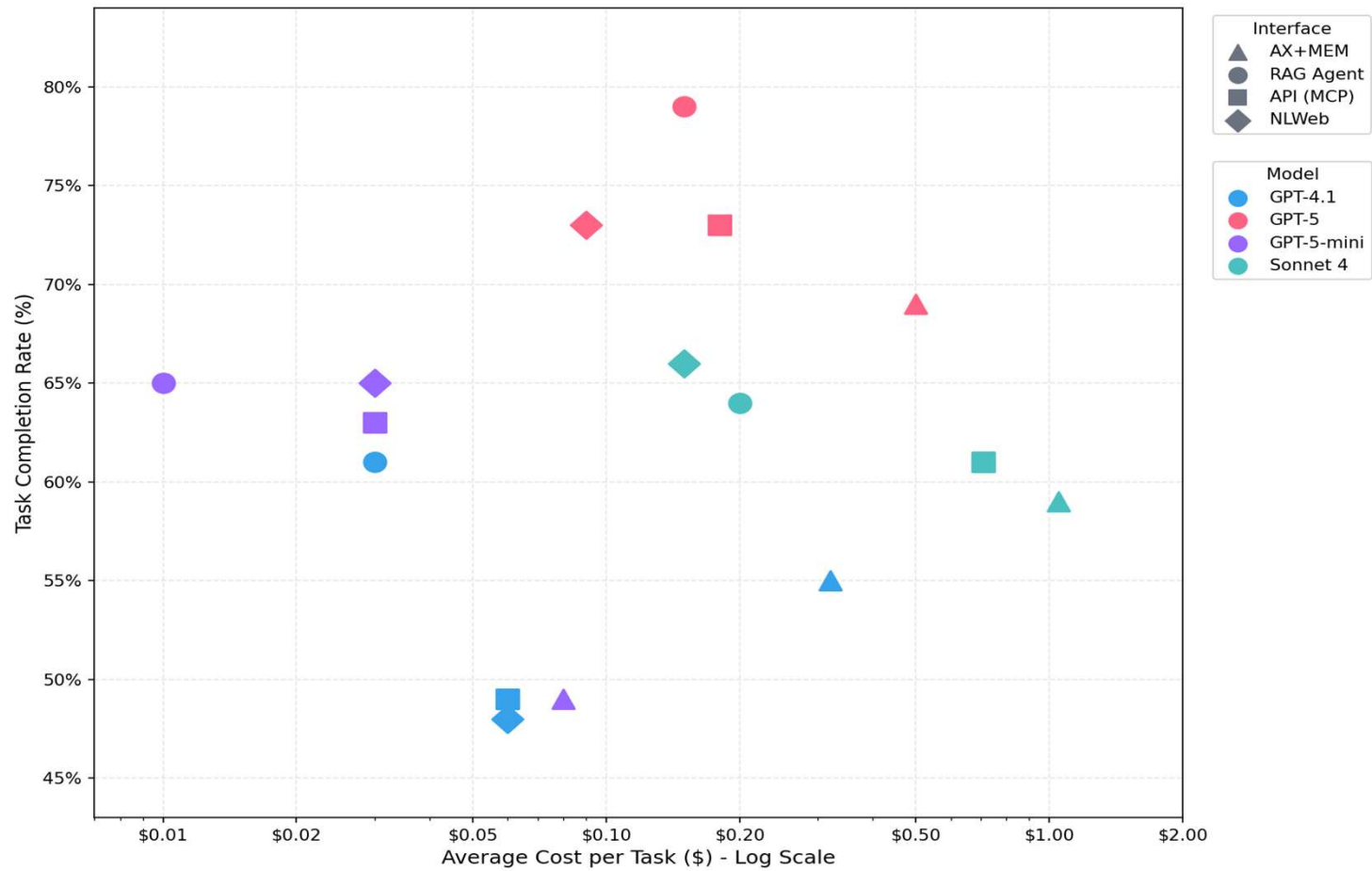
- Token usage differs greatly
- RAG consumes on average 1/5 of the tokens compared to HTML
- NLWeb consumes 1/3 the tokens of HTML

Agent	CR	F1	Token	Cost	Runtime
HTML	0.57	0.67	241,136	\$0.52	291 s
RAG	0.68	0.77	46,667	\$0.10	50 s
MCP	0.62	0.75	139,569	<u>\$0.27</u>	62 s
NLWeb	<u>0.64</u>	<u>0.76</u>	<u>71,214</u>	\$0.10	<u>53 s</u>

* Results are averaged across all tasks and models and aggregated by interface.

→ RAG is the most token efficient

Price to performance



Steiner - MCP vs RAG vs NLWeb vs HTML - 02.12.2025

Summary & Practical Guidance

- RAG, MCP, and NLWeb outperform HTML by **≈9 F1 points** on average.
- The largest gains appear in **well-specified search tasks**.
- All interfaces show **lower performance** on vague and cheapest-product tasks.
- **RAG and NLWeb** are the most efficient, using **2–5× fewer tokens** and running significantly faster.
- Efficiency comes mainly from **reduced input tokens** and avoiding HTML navigation.
- **RAG and NLWeb** deliver the highest effectiveness and efficiency.

Thank you and Questions



Paper



Website



LinkedIn

Backup: Average F1 performance by task set.
Best bolded, 2nd best underlined.

Task set	HTML	RAG	MCP	NLWeb
Specific Product Search	0.77	<u>0.91</u>	0.90	0.92
Vague Product Search	<u>0.63</u>	0.61	0.59	0.66
Cheapest Product Search	0.61	0.68	<u>0.63</u>	0.60
Action & Transaction	0.77	0.86	0.92	<u>0.91</u>

Backup: Average F1 performance by task set.
Best bolded, 2nd best underlined.

Task set	HTML	RAG	MCP	NLWeb
Specific Product Search	0.77	<u>0.91</u>	0.90	0.92
Vague Product Search	<u>0.63</u>	0.61	0.59	0.66
Cheapest Product Search	0.61	0.68	<u>0.63</u>	0.60
Action & Transaction	0.77	0.86	0.92	<u>0.91</u>

Results for Specific Product Search. Best bolded, 2nd best underlined.

Agent	Model	CR	F1	Token	Cost	Runtime
HTML	GPT-4.1	0.52	0.74	148,833	\$0.32	92 s
HTML	GPT-5	0.74	0.83	314,231	\$0.62	624 s
HTML	GPT-5-mini	0.57	0.79	233,453	\$0.09	236 s
HTML	Sonnet 4	0.61	0.72	327,640	\$1.23	357 s
RAG	GPT-4.1	0.74	0.86	12,534	<u>\$0.03</u>	7 s
RAG	GPT-5	<u>0.83</u>	0.96	100,707	\$0.18	123 s
RAG	GPT-5-mini	0.78	<u>0.93</u>	32,565	\$0.01	47 s
RAG	Sonnet 4	<u>0.83</u>	0.91	44,527	\$0.15	24 s
MCP	GPT-4.1	0.74	0.88	<u>24,190</u>	\$0.05	<u>10 s</u>
MCP	GPT-5	0.87	0.96	134,190	\$0.20	97 s
MCP	GPT-5-mini	0.74	0.90	90,948	<u>\$0.03</u>	80 s
MCP	Sonnet 4	0.65	0.84	244,762	\$0.75	41 s
NLWeb	GPT-4.1	0.57	0.84	26,665	\$0.06	13 s
NLWeb	GPT-5	0.87	0.96	42,380	\$0.07	62 s
NLWeb	GPT-5-mini	0.78	<u>0.93</u>	97,861	<u>\$0.03</u>	105 s
NLWeb	Sonnet 4	<u>0.83</u>	0.92	37,005	\$0.12	15 s

Results for Vague Product Search. Best bolded, 2nd best underlined.

Agent	Model	CR	F1	Token	Cost	Runtime
HTML	GPT-4.1	0.32	0.49	157,639	\$0.34	97 s
HTML	GPT-5	0.60	<u>0.78</u>	288,320	\$0.60	641 s
HTML	GPT-5-mini	0.32	0.60	255,852	\$0.09	248 s
HTML	Sonnet 4	0.42	0.60	337,967	\$1.25	361 s
RAG	GPT-4.1	0.26	0.53	18,198	\$0.04	9 s
RAG	GPT-5	0.74	0.82	111,872	\$0.20	147 s
RAG	GPT-5-mini	0.58	0.66	68,452	\$0.02	83 s
RAG	Sonnet 4	0.37	0.41	92,409	\$0.30	42 s
MCP	GPT-4.1	0.11	0.46	27,456	\$0.06	<u>12 s</u>
MCP	GPT-5	0.47	0.65	154,716	\$0.23	119 s
MCP	GPT-5-mini	0.53	0.73	105,082	<u>\$0.03</u>	101 s
MCP	Sonnet 4	0.32	0.53	276,672	\$0.85	53 s
NLWeb	GPT-4.1	0.37	0.60	<u>27,689</u>	\$0.06	13 s
NLWeb	GPT-5	0.53	0.72	77,641	\$0.12	62 s
NLWeb	GPT-5-mini	<u>0.63</u>	0.77	119,435	\$0.04	126 s
NLWeb	Sonnet 4	0.37	0.55	62,435	\$0.20	26 s

Results for Cheapest Product Search. Best bolded, 2nd best underlined.

Agent	Model	CR	F1	Token	Cost	Runtime
HTML	GPT-4.1	0.50	0.58	149,657	\$0.32	97 s
HTML	GPT-5	0.75	0.75	194,594	\$0.37	465 s
HTML	GPT-5-mini	0.54	0.59	209,629	\$0.08	197 s
HTML	Sonnet 4	0.54	0.54	256,543	\$0.93	264 s
RAG	GPT-4.1	0.62	0.68	16,958	\$0.04	9 s
RAG	GPT-5	<u>0.72</u>	0.78	70,736	\$0.14	129 s
RAG	GPT-5-mini	0.69	<u>0.76</u>	29,754	\$0.01	46 s
RAG	Sonnet 4	0.50	0.50	70,887	\$0.23	35 s
MCP	GPT-4.1	0.42	0.60	27,816	\$0.06	9 s
MCP	GPT-5	0.69	0.72	94,017	\$0.14	76 s
MCP	GPT-5-mini	0.54	0.57	95,723	<u>\$0.03</u>	75 s
MCP	Sonnet 4	0.62	0.63	217,687	\$0.67	41 s
NLWeb	GPT-4.1	0.27	0.42	<u>25,941</u>	\$0.05	<u>13 s</u>
NLWeb	GPT-5	0.65	0.75	58,185	\$0.09	60 s
NLWeb	GPT-5-mini	0.54	0.59	71,877	<u>\$0.03</u>	93 s
NLWeb	Sonnet 4	0.58	0.63	42,790	\$0.14	21 s

Results for Action & Transaction. Best bolded, 2nd best underlined.

Agent	Model	CR	F1	Token	Cost	Runtime
HTML	GPT-4.1	1.00	1.00	124,782	\$0.27	79 s
HTML	GPT-5	0.67	0.64	182,320	\$0.35	395 s
HTML	GPT-5-mini	0.53	0.56	149,165	\$0.05	153 s
HTML	Sonnet 4	0.87	0.87	196,076	\$0.71	189 s
RAG	GPT-4.1	0.87	0.96	8,445	<u>\$0.02</u>	6 s
RAG	GPT-5	<u>0.93</u>	<u>0.98</u>	18,631	\$0.04	35 s
RAG	GPT-5-mini	0.47	0.54	<u>12,380</u>	\$0.01	26 s
RAG	Sonnet 4	<u>0.93</u>	<u>0.98</u>	17,635	\$0.06	27 s
MCP	GPT-4.1	0.73	0.86	45,714	\$0.09	<u>13 s</u>
MCP	GPT-5	<u>0.93</u>	<u>0.98</u>	193,062	\$0.28	88 s
MCP	GPT-5-mini	0.67	0.86	289,448	\$0.08	160 s
MCP	Sonnet 4	<u>0.93</u>	<u>0.98</u>	182,728	\$0.56	40 s
NLWeb	GPT-4.1	0.87	0.96	38,565	\$0.08	14 s
NLWeb	GPT-5	<u>0.93</u>	<u>0.98</u>	83,333	\$0.14	50 s
NLWeb	GPT-5-mini	0.53	0.74	174,717	\$0.05	108 s
NLWeb	Sonnet 4	<u>0.93</u>	<u>0.98</u>	46,833	\$0.15	19 s

Efficiency overview by agent and model.

Interface	Model	Token	Cost	Runtime
HTML	Average	225,090	\$0.49	281 s
HTML	GPT-4.1	146,761	\$0.32	92 s
HTML	GPT-5	253,759	\$0.50	522 s
HTML	GPT-5-mini	215,885	\$0.08	211 s
HTML	Sonnet 4	283,956	\$1.05	299 s
RAG	Average	47,093	\$0.10	51 s
RAG	GPT-4.1	14,477	\$0.03	8 s
RAG	GPT-5	79,142	\$0.15	114 s
RAG	GPT-5-mini	36,252	\$0.01	51 s
RAG	Sonnet 4	58,885	\$0.20	32 s
MCP	Average	121,624	\$0.25	57 s
MCP	GPT-4.1	29,964	\$0.06	<u>11 s</u>
MCP	GPT-5	119,841	\$0.18	94 s
MCP	GPT-5-mini	104,319	\$0.03	80 s
MCP	Sonnet 4	232,374	\$0.71	44 s
NLWeb	Average	57,840	\$0.08	49 s
NLWeb	GPT-4.1	<u>28,876</u>	\$0.06	14 s
NLWeb	GPT-5	56,922	\$0.09	59 s
NLWeb	GPT-5-mini	98,449	\$0.03	102 s
NLWeb	Sonnet 4	46,415	\$0.15	20 s