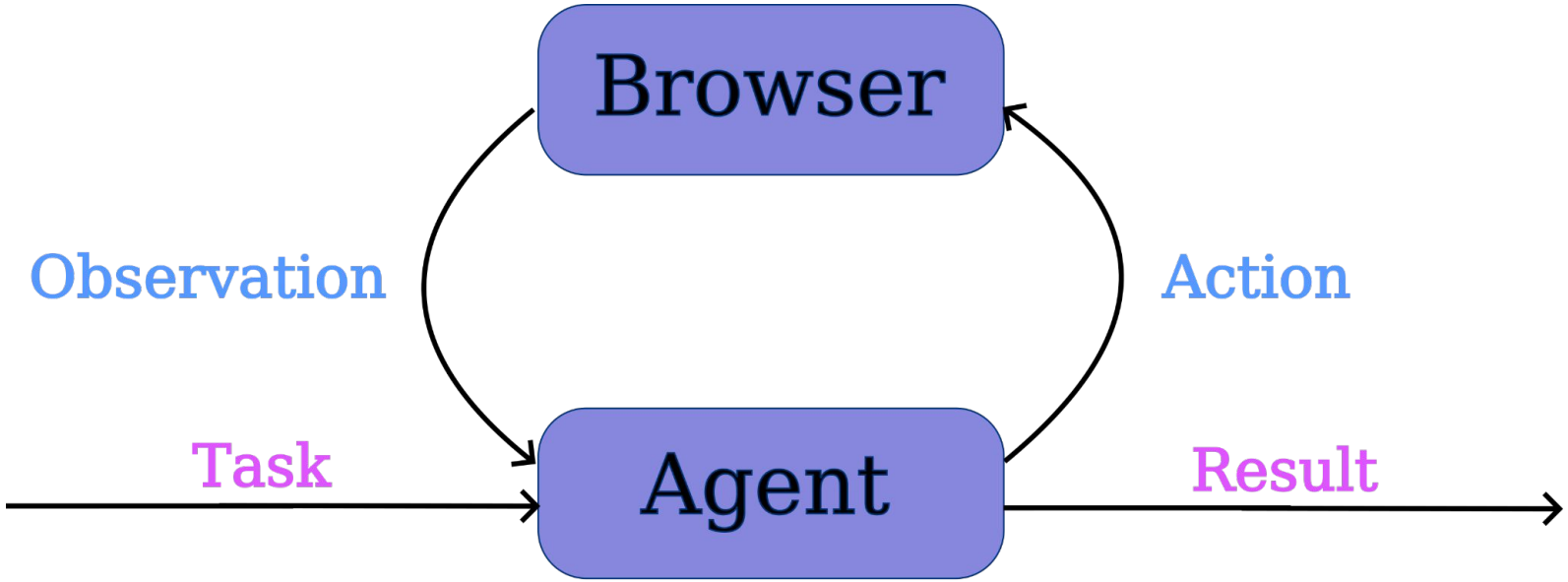
MUNI
FI

WebMap: Improving LLM Web Agents with Semantic Search for Relevant Web Pages

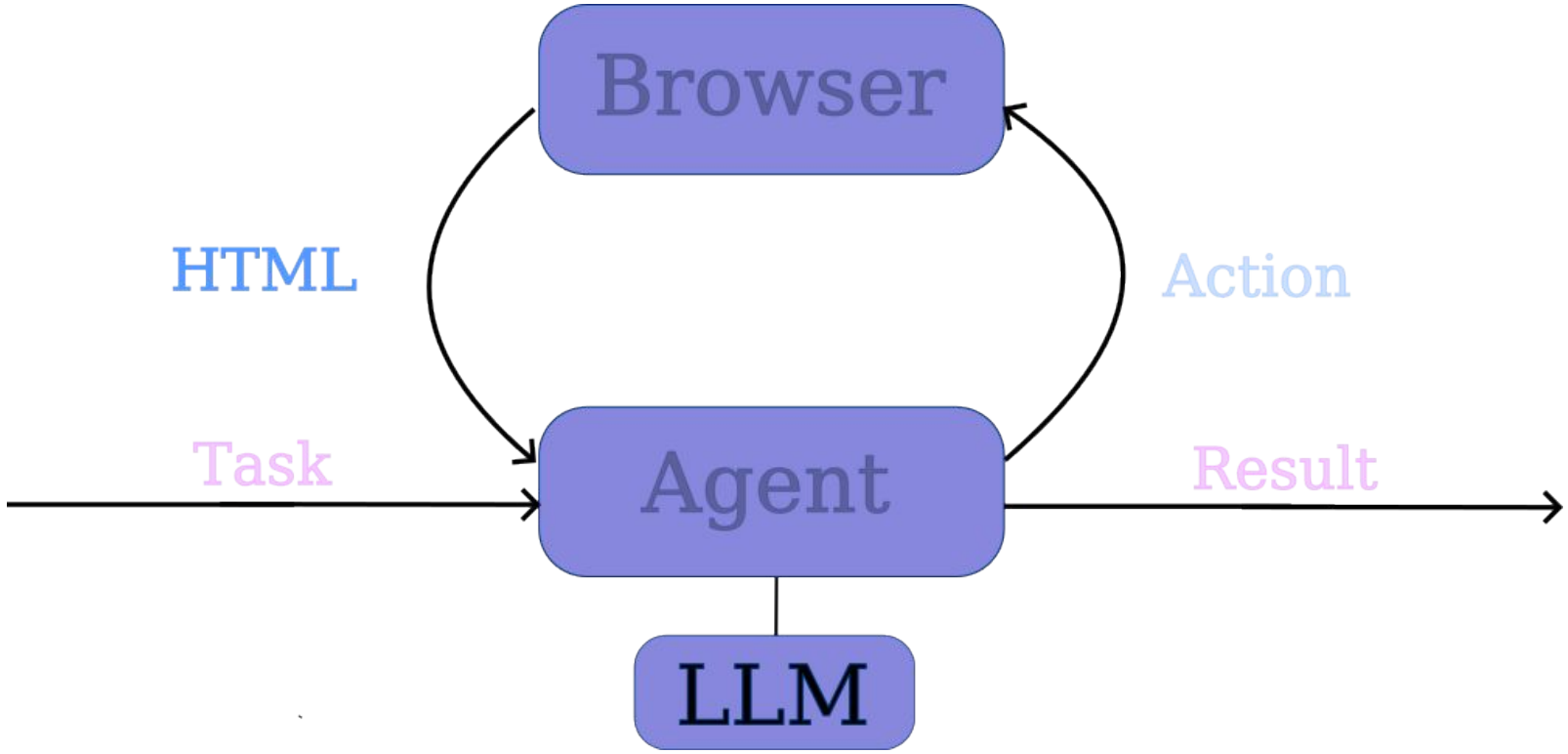
Michal Spiegel and Aleš Horák

Task Description



MUNI
FI

Task Description

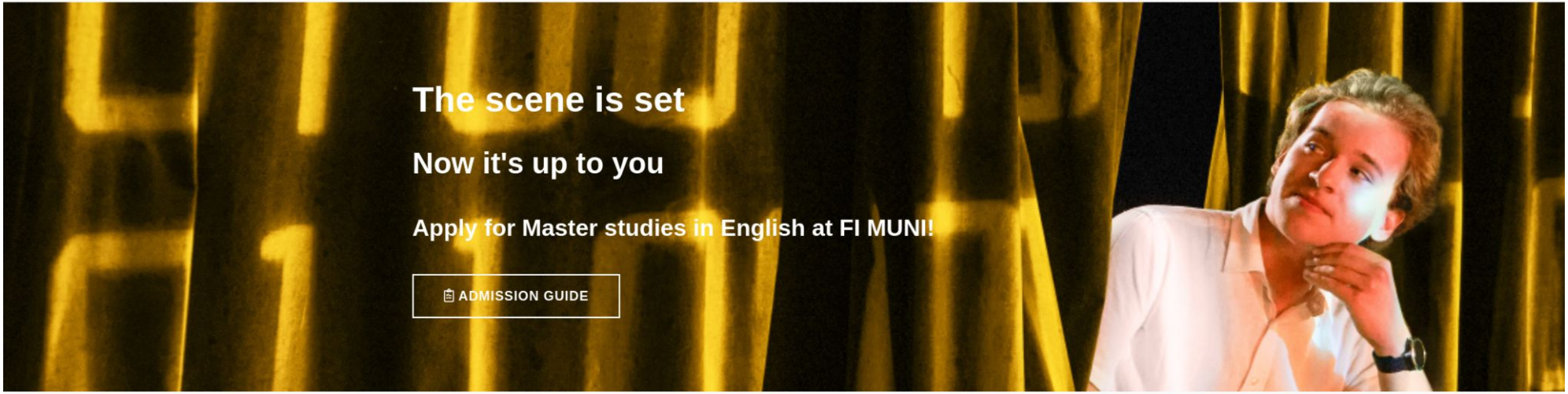


MUNI
FI

Example Task

Task description: **Find the phone number of the study department**

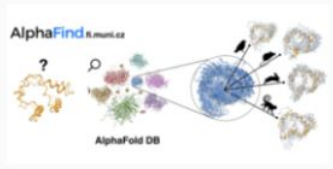
MUNI
FI



NEWS FROM FI

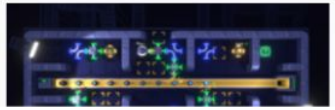
FI MU student attends a prestigious AI conference in Canada

FI MU teams shine in the NXP Cup finals in Hamburg



AlphaFind: a new compass in the world of proteins

Computer Science Students Present Generative Art at Mind The Passage Exhibition



```
<html class="js no-touch" style="" lang="cs">
<head>
	<title>FI MU</title>
	<meta http-equiv="Content-Type" content="text/html; charset=utf-8">

	<link rel="canonical" href="https://www.fi.muni.cz/index.html.cs">

<header class="header">
	<div class="row-main">
	<div class="header__wrap">
	<div class="header__side">

		<!-- VYHLEDAVANI -->

		<form id="searchbox_010477525560646045864:huk3as1vcri" action="//www.fi.muni.cz/app/search" class="header__search" role="search">
			<fieldset>
			<input type="hidden" name="cx" value="010477525560646045864:huk3as1vcri">
			<input type="hidden" name="cof" value="FORID:9">
			<input type="hidden" name="ie" value="UTF-8">
			<input type="hidden" name="oe" value="UTF-8">
			<input type="hidden" name="lang" value="cs">
			<p>
				<label for="search" class="header__search__label icon icon-search">
				<span class="vhide">Vyhledávání</span>
				</label>
				<span class="inp-fix inp-icon inp-icon--after">
				<input name="q" id="search" class="inp-text" placeholder="Hledat..." type="text">
				<button type="submit" class="btn-icon icon icon-search">
				<span class="vhide">Vyhledat</span>
				</button>
				</span>
```

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
    <title>FI MU</title>
…
…

Generated output:

I will issue the action to click the element with ID 45 and the text ‘Contacts’. Action: click(“45”).

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
<title>FI MU</title>
…
…

Generated output:

I will issue the action to click the element with ID 45 and the text ‘Contacts’. Action: click(“45”).

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
    <title>FI MU</title>
…
…

Generated output:

I will issue the action to click the element with ID 45 and the text ‘Contacts’. **Action: click(“45”).**

ABOUT US

- About us
- People
- Career
- Vacancies
- Ethics and protection of rights
- Departments
- History
- Awards and honors
- Documents
- Events and conferences
- Event Photo Gallery
- Contacts

ADMISSIONS

- Programmes offered
- Why FI
- Admission to English study
- Admission to Czech study
- Application
- Information for students
- Advanced master's state examination
- Doctoral study
- Lifelong learning
- Open days

RESEARCH

- Scientific research and development
- Areas of research
- Research groups
- Doctoral study
- Publications
- Lectures and Informatics colloquium
- Scientific Board
- Administration and support of R&D

COOPERATION

- Cooperation possibilities
- Association of Industrial Partners
- Secondary school activities
- CERIT Science Park
- IS MU outsourcing
- Cooperation offers & Fundraising

click

IS INET MU Tech info FAdmin Contacts Home

Contacts

Address	Botanická 68a 602 00 Brno Czech Republic	**IČ**	00216224
Telephone	+420-549 49 1810	**DIČ**	CZ00216224
Fax	+420-549 49 1820	**IBAN**	CZ43 0100 0000 0000 8563 6621
E-mail	info@fi.muni.cz	**SWIFT**	KOMBCZPPXXX

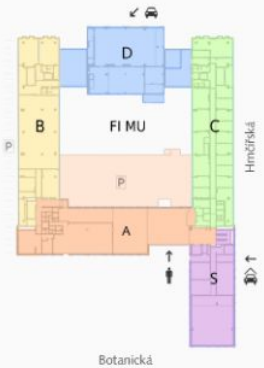
Office for Studies

Address:
Office for Studies at FI
Botanická 68a
602 00 Brno

Location in the building:
building A, 2nd floor,
rooms A202, A203, A204, A205, A206

E-mail	studijni@fi.muni.cz (preferred contact)
Telephone	+420-549 49 1818
Fax	+420-549 49 1820

Map of the faculty buildings



3D TOUR

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
<title>Contacts</title>
…
…

Generated output:

The previous action successfuly moved us to the Contacts page. I can now see a section describing the Office for Studies. The phone number of the study department is +420-549 49 1818. Action: finish(“+420-549 49 1818”).

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
<title>Contacts</title>
…
…

Generated output:

The previous action successfuly moved us to the Contacts page. I can now see a section describing the Office for Studies. The phone number of the study department is +420-549 49 1818. Action: finish(“+420-549 49 1818”).

LLM Prompt:

You are a helpful assistant that can assist with web navigation tasks.
You are given a simplified html webpage and a task description.
Your goal is to complete the task. You can use the provided functions below to interact with the current webpage.

#Provided functions:
click(element_id: str) -> None:
hover(element_id: str) -> None:
select(element_id: str, option: str) -> None:
finish(answer: Optional[str]) -> None:

#Task: Find the phone number of the study department

#Observation:
<html>
<head>
    <title>Contacts</title>
…
…

Generated output:

The previous action successfuly moved us to the Contacts page. I can now see a section describing the Office for Studies. The phone number of the study department is +420-549 49 1818. **Action: finish(“+420-549 49 1818”).**

MUNI
FI

Interface Learning Curves

MUNI FI

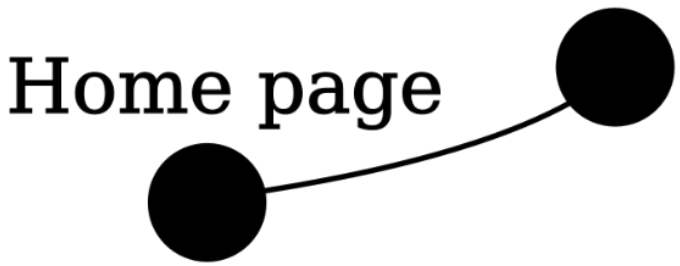
Universal Interface

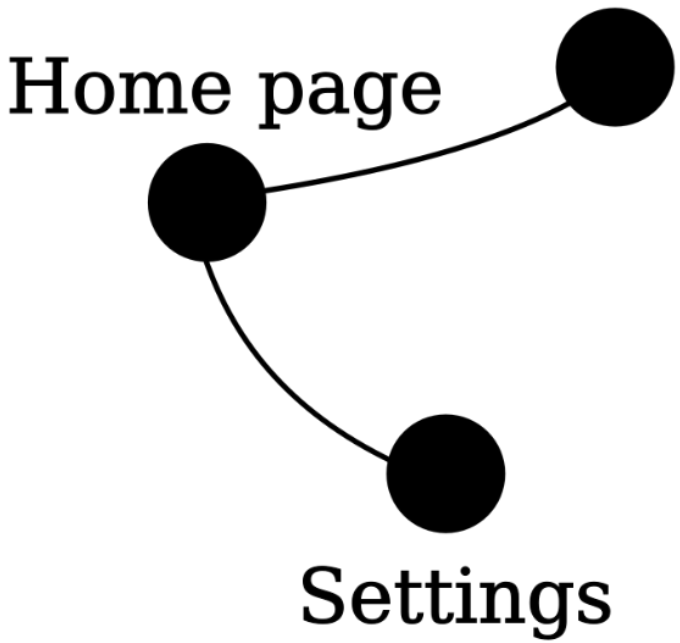
MUNI FI

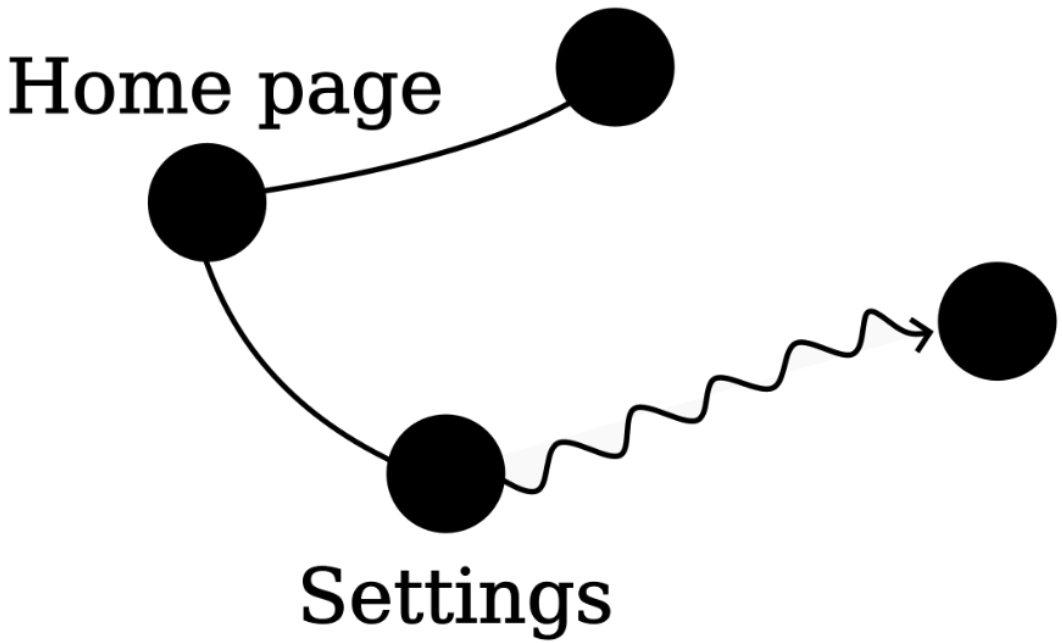
What's the problem?

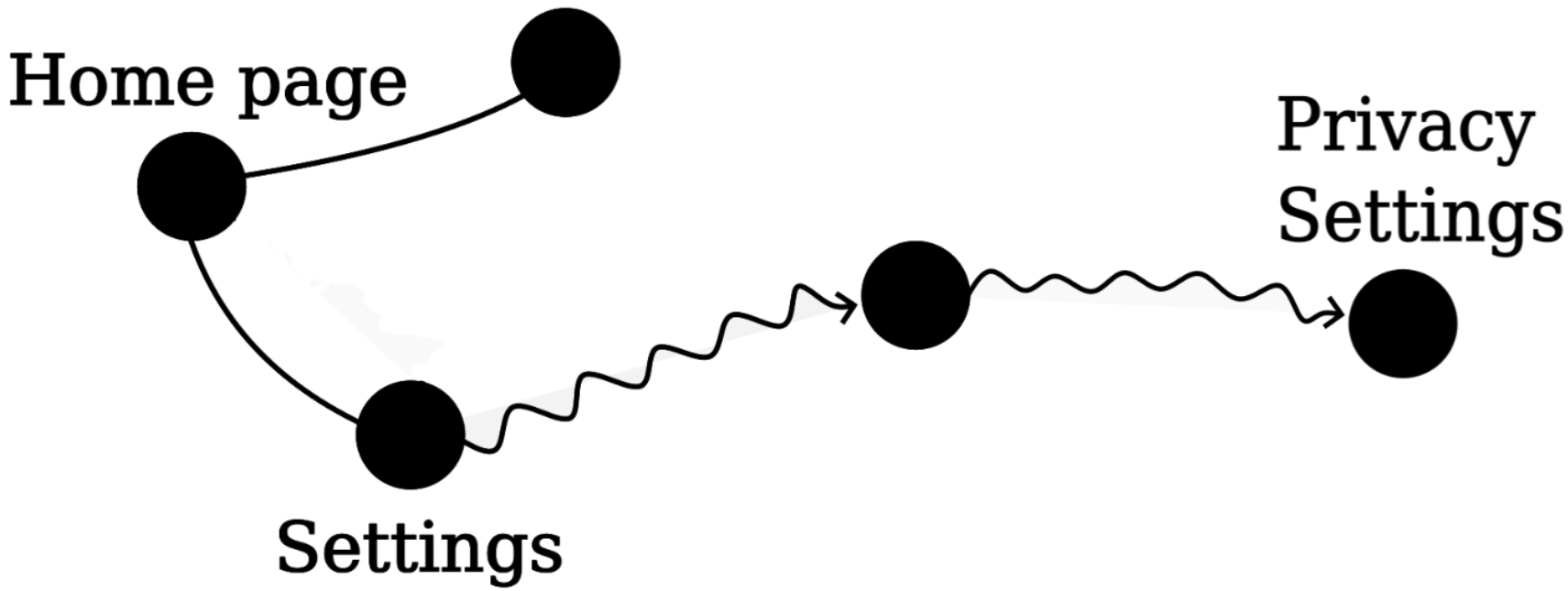
MUNI
FI

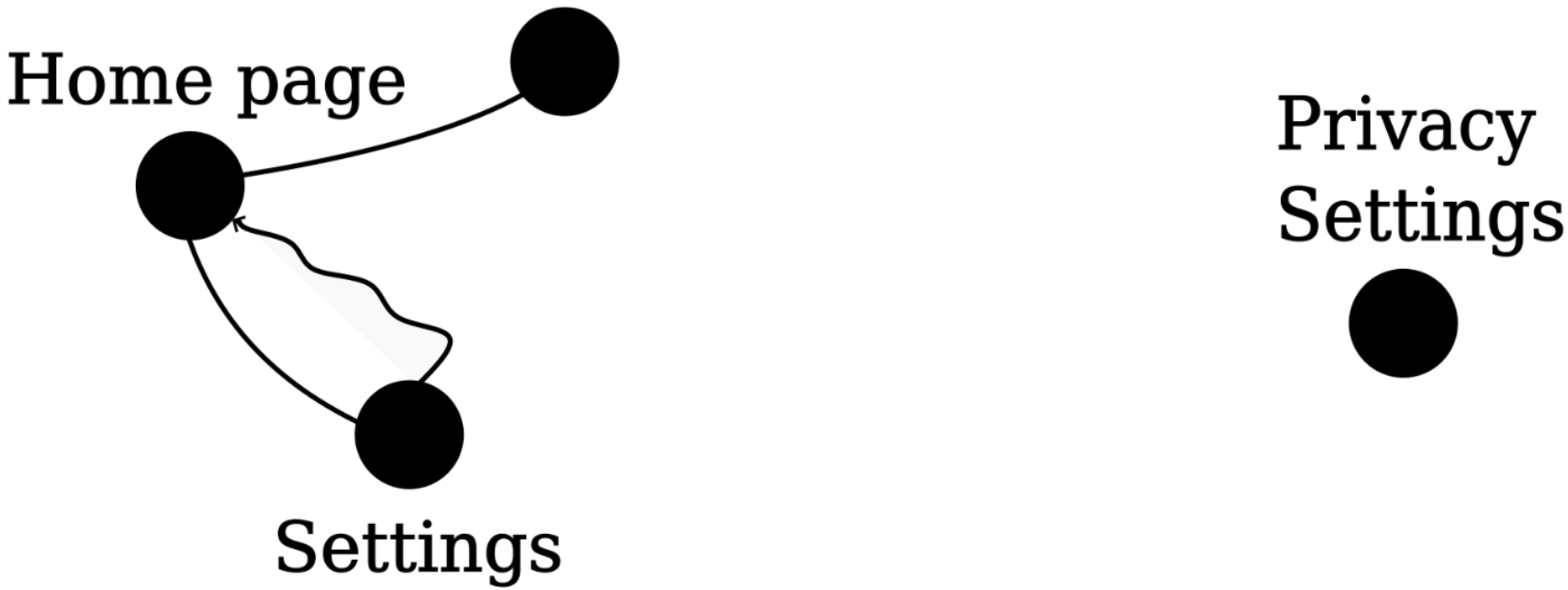
Home page


Home page


Home page
Settings


Home page
Settings

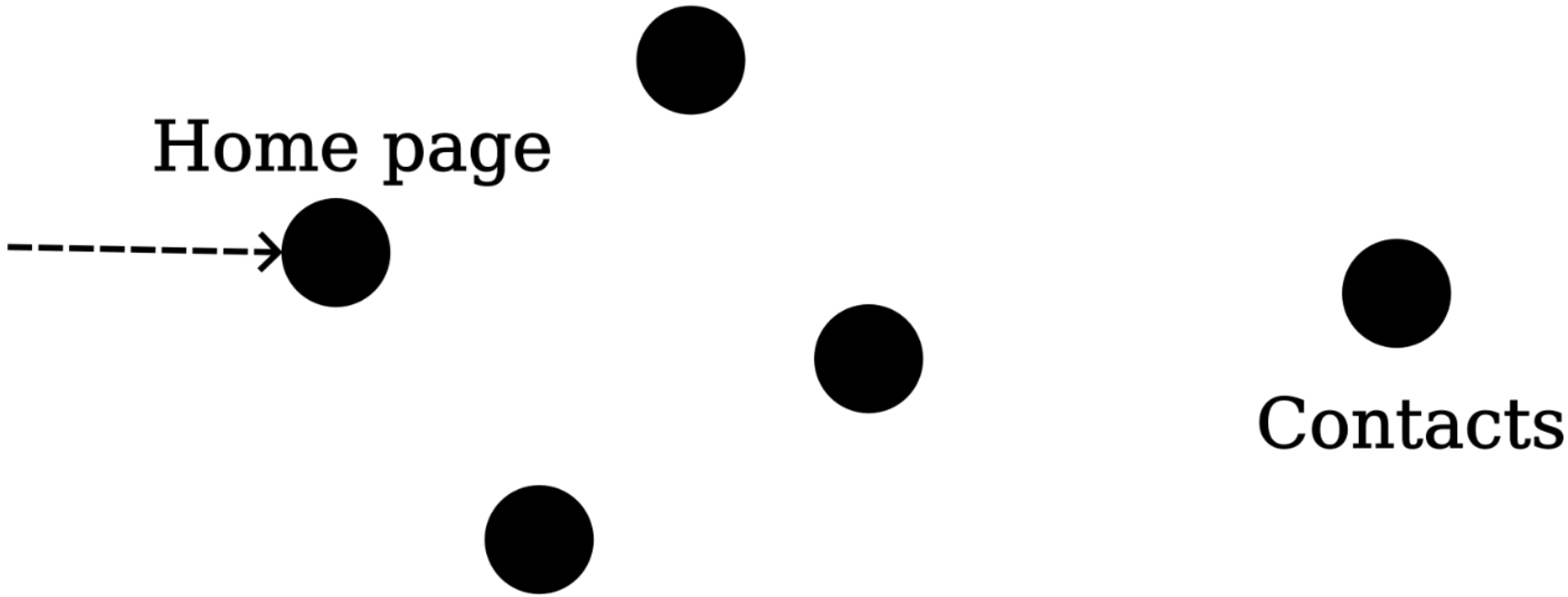

Home page
Settings
Privacy
Settings


Home page
Settings
Privacy
Settings

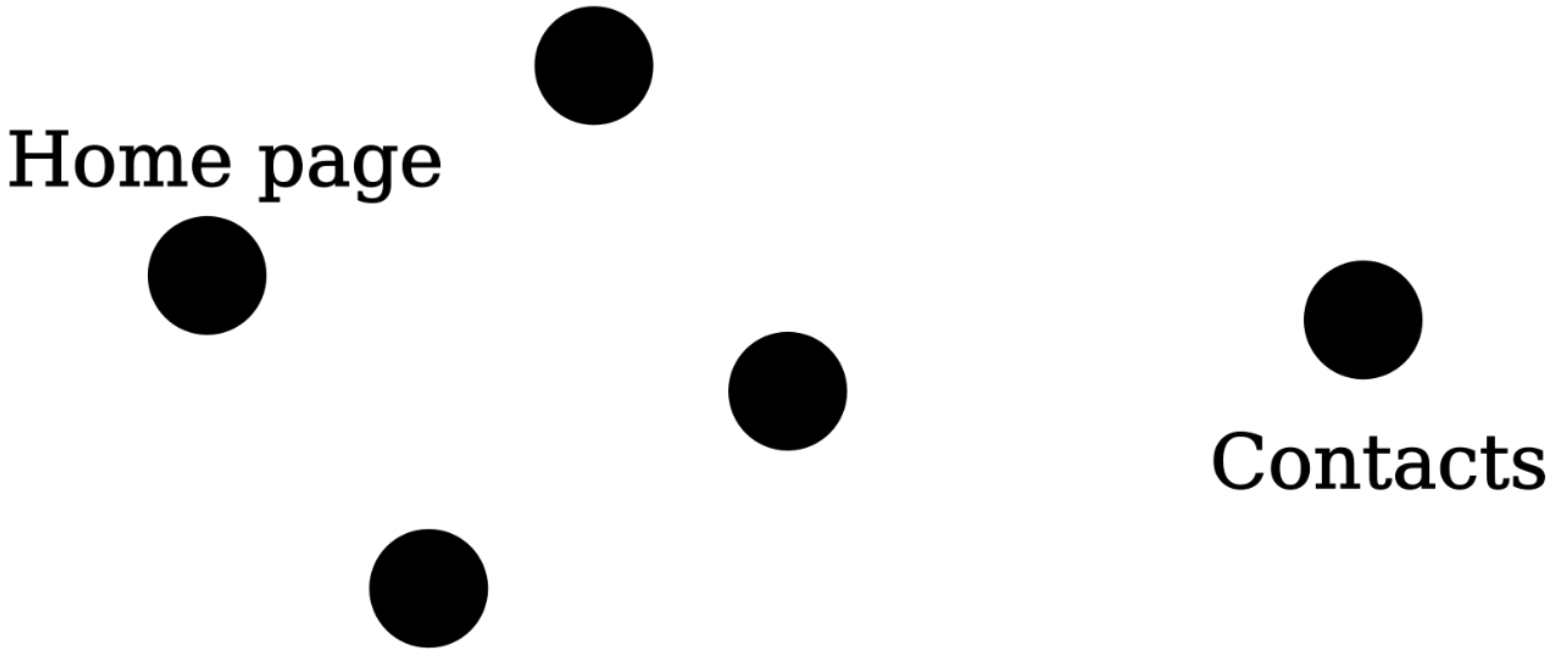
Type of Error	Count	Percentage
Controls Understanding	24	30
Incorrect Web Page	18	23
Observation Understanding	16	20
Reasoning Error	10	13
Planning Error	5	6
Hallucinations	4	5
Evaluation Dataset Fault	1	1
Not enough information in the observation	1	1
Fail to recover from an error	1	1

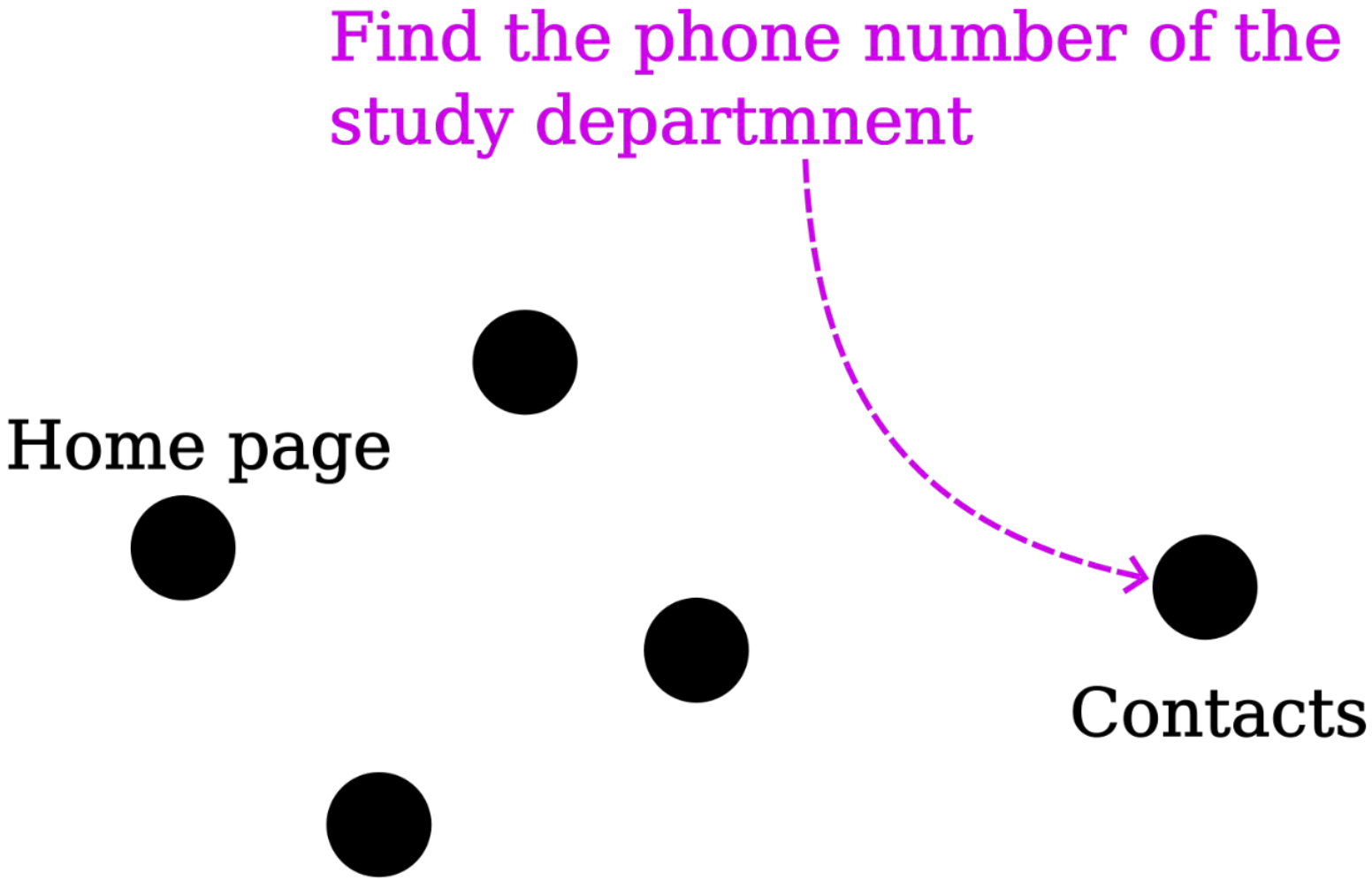
Proposed Solution

MUNI
FI


Home page
Contacts

Find the phone number of the study departmnent




Find the phone number of the study departmnent
Home page
Contacts

Key Contributions

- **Design and implementation of WebMap**, a novel method that improves the accuracy and efficiency of autonomous agents on the web

MUNI
FI

Key Contributions

- Design and implementation of WebMap, a novel method that improves the accuracy and efficiency of autonomous agents on the web
- **Experimental evaluation and analysis** of the proposed method on the WebArena benchmark

MUNI
FI

Key Contributions

- Design and implementation of WebMap, a novel method that improves the accuracy and efficiency of autonomous agents on the web
- Experimental evaluation and analysis of the proposed method on the WebArena benchmark
- **Error analysis** of WebMap and baseline approaches on the WebArena benchmark, highlighting important challenges for future work

MUNI
FI

Key Contributions

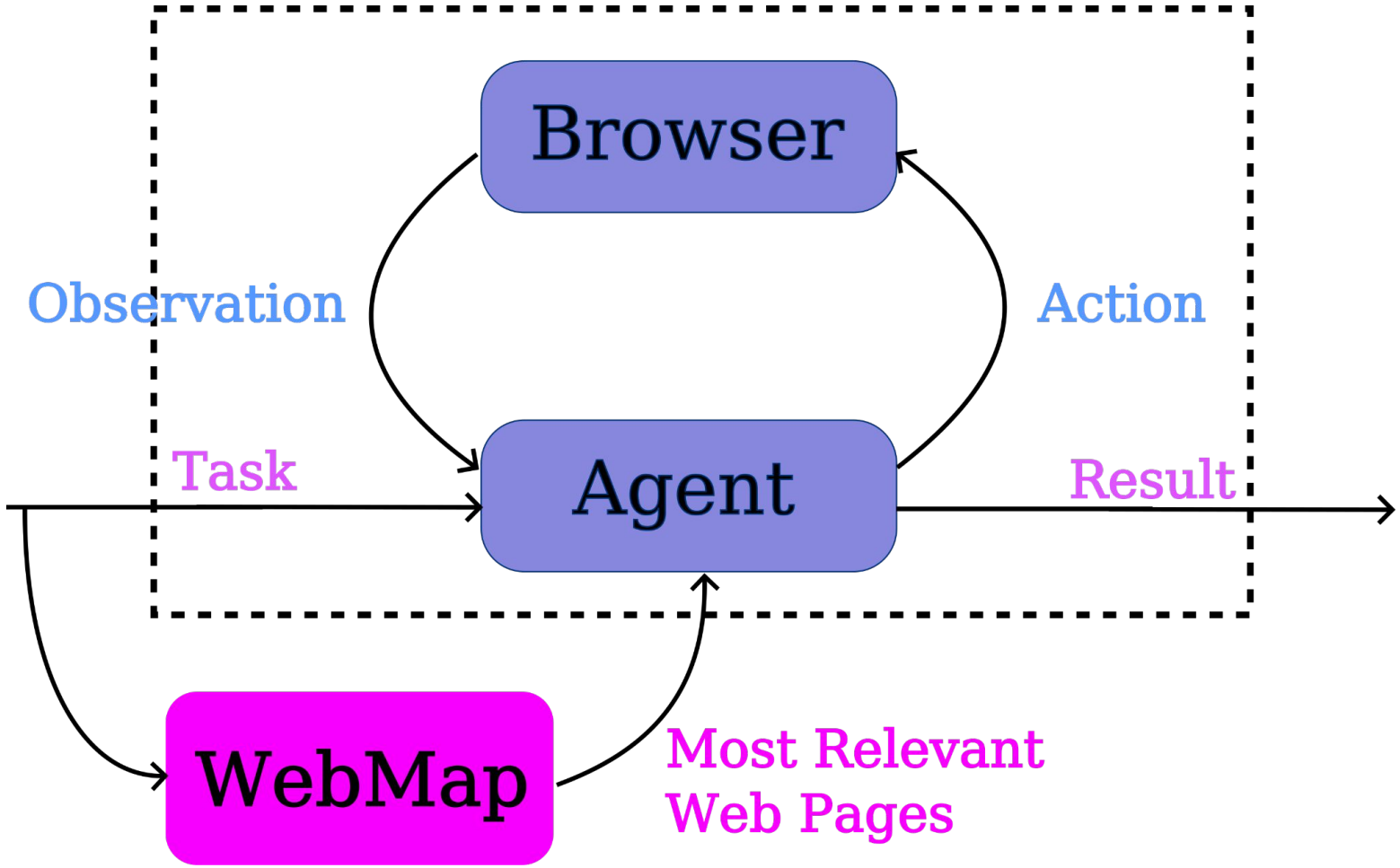
- **Design and implementation of WebMap**, a novel method that improves the accuracy and efficiency of autonomous agents on the web
- **Experimental evaluation and analysis** of the proposed method on the WebArena benchmark
- **Error analysis** of WebMap and baseline approaches on the WebArena benchmark, highlighting important challenges for future work

MUNI
FI

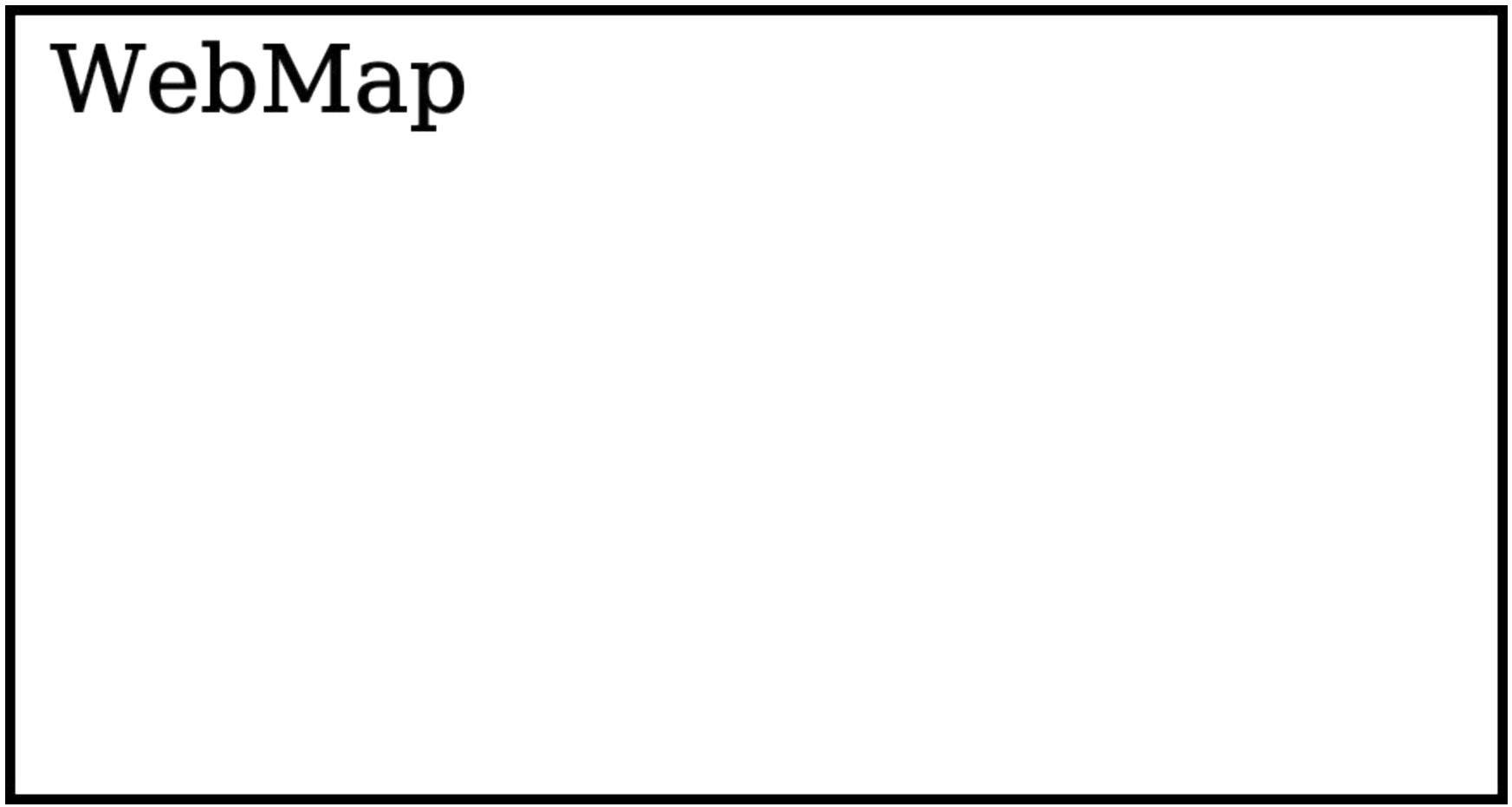
Designing WebMap

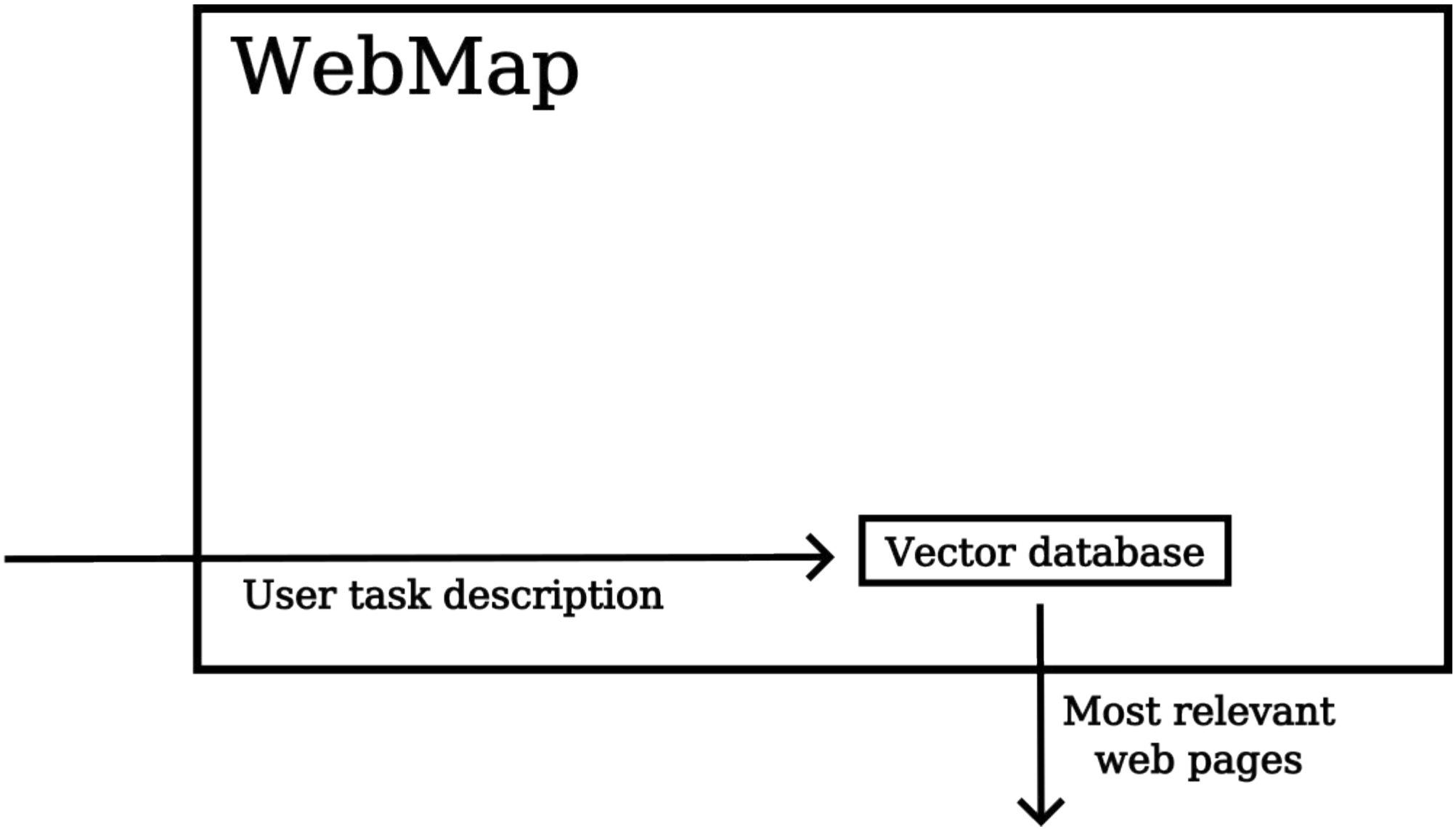
MUNI
FI

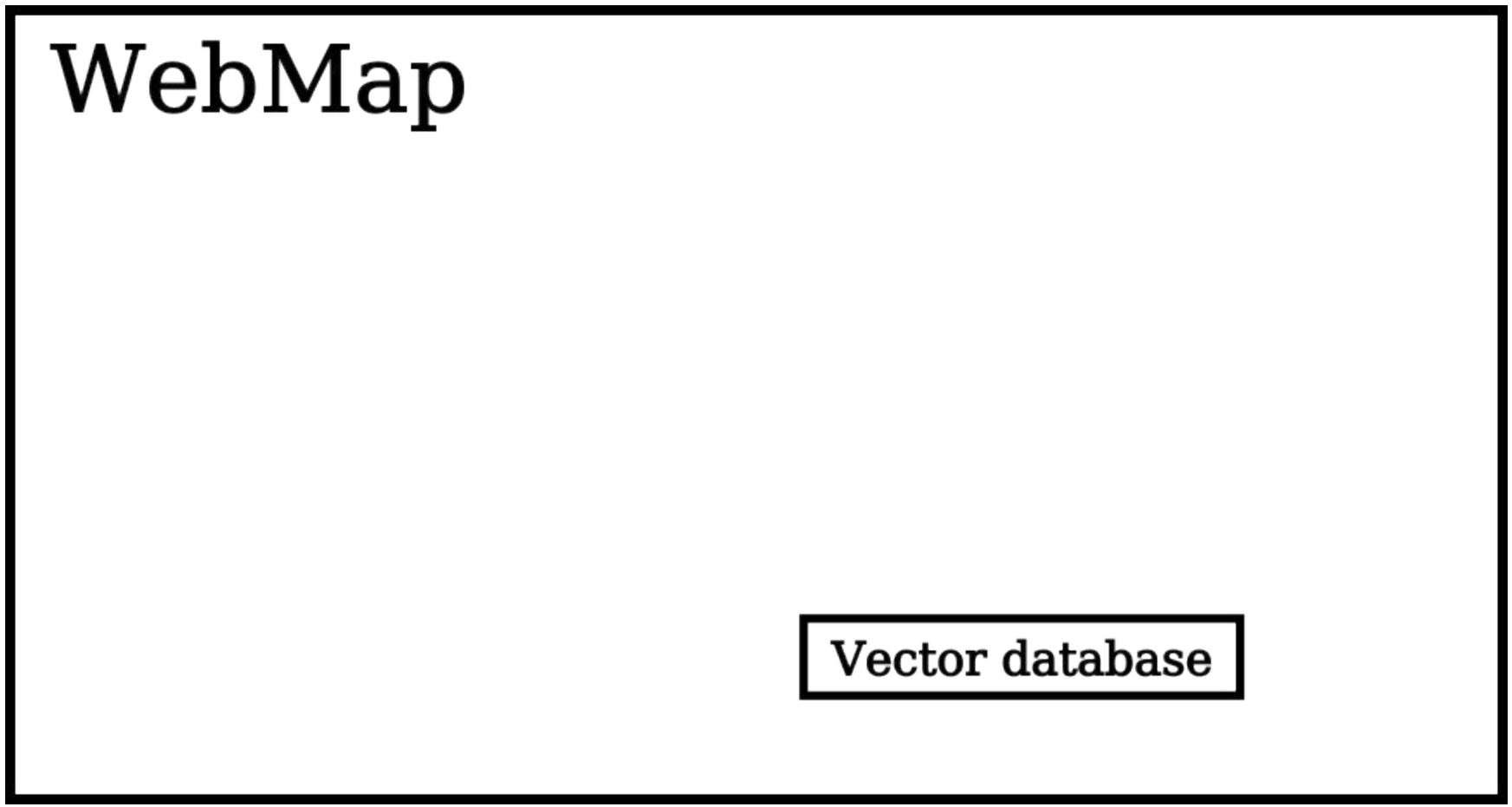
Designing WebMap



MUNI
FI


WebMap

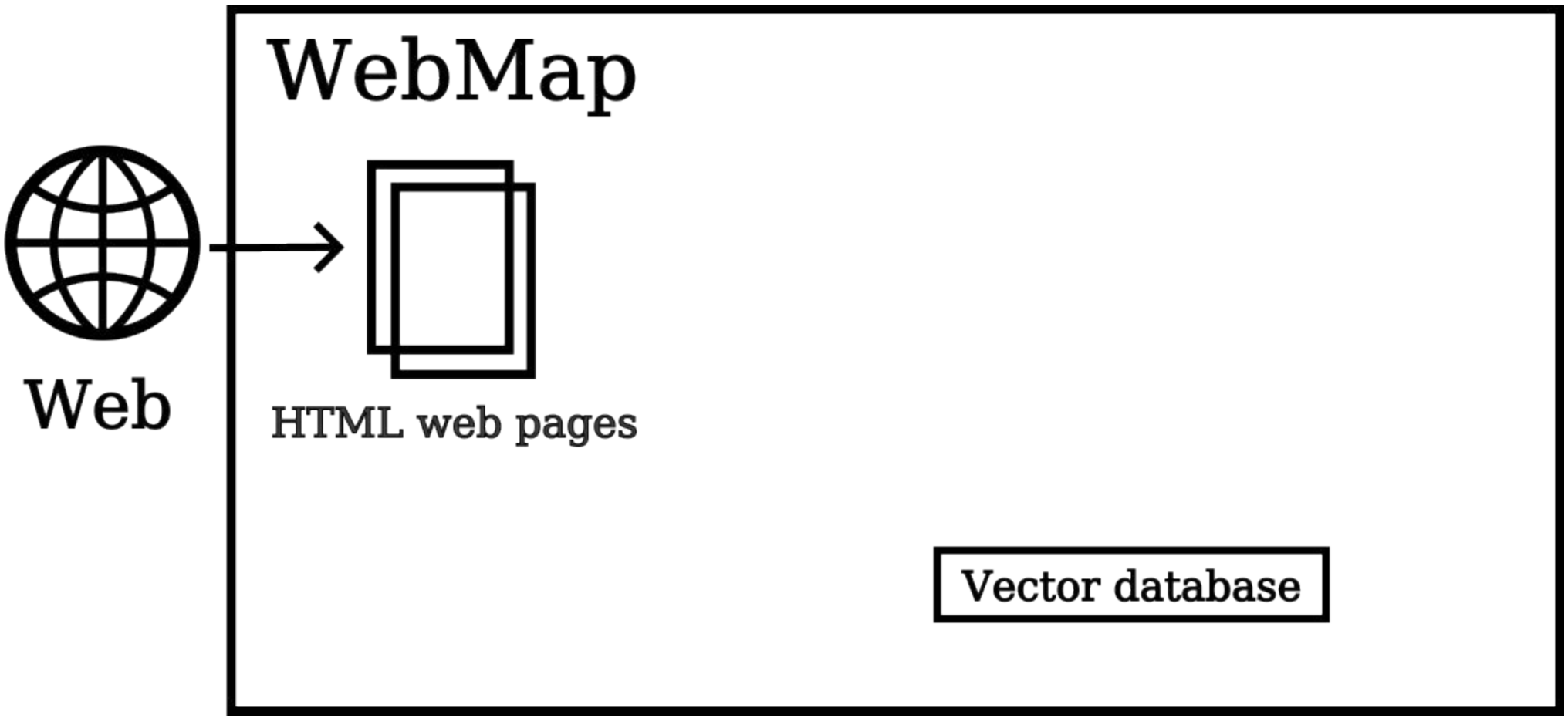

WebMap
Vector database
User task description
Most relevant
web pages

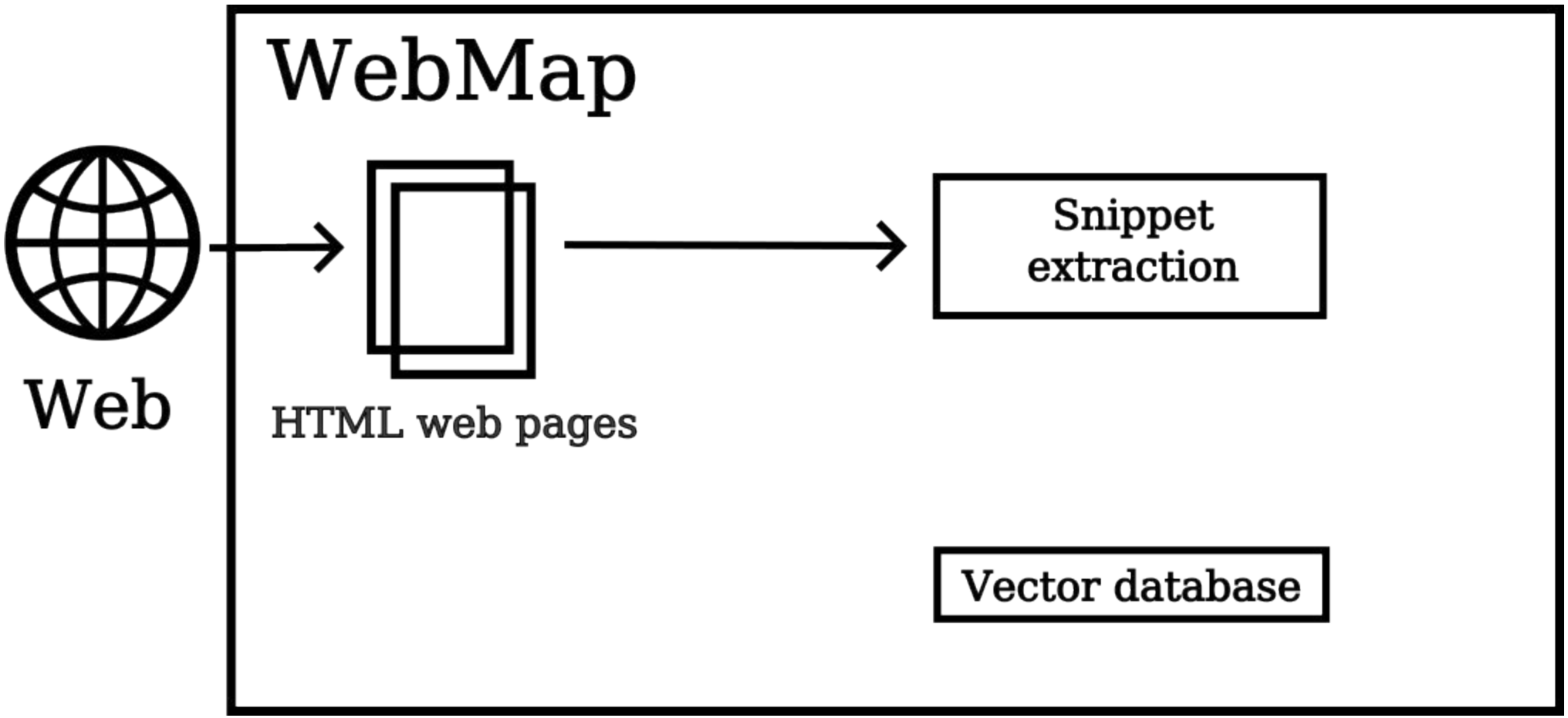

WebMap
Vector database

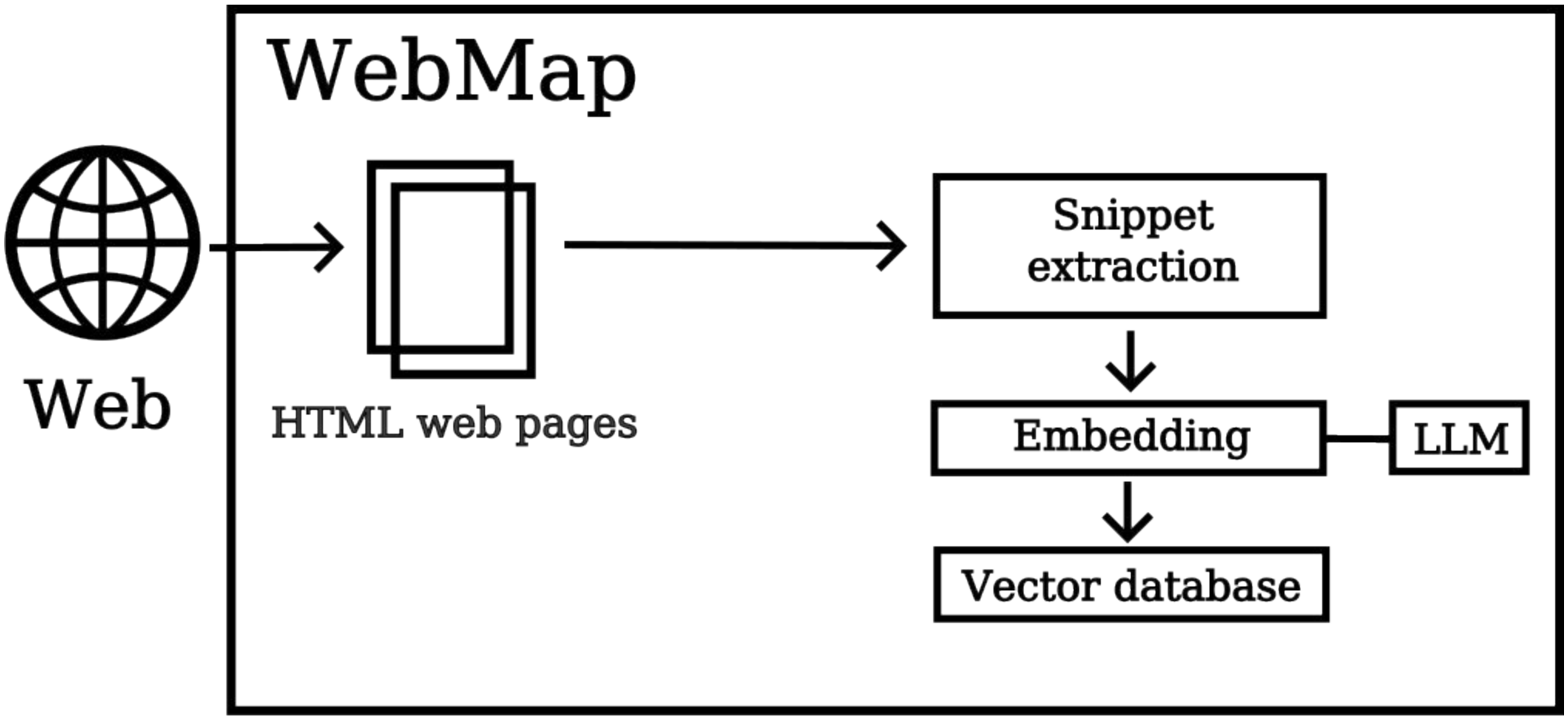
Web

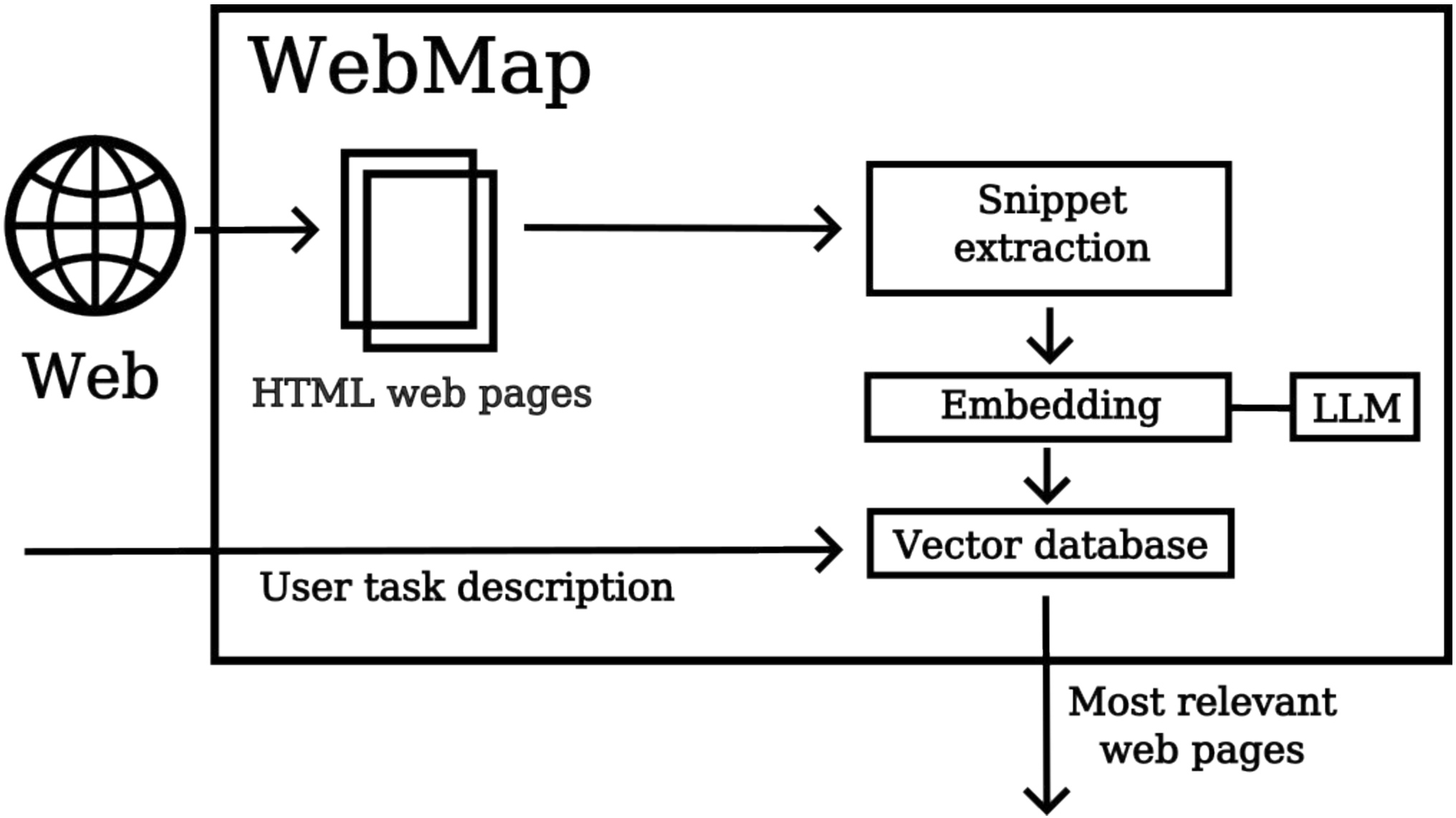
WebMap

Vector database


WebMap
Web
HTML web pages
Vector database


WebMap
Web
HTML web pages
Snippet extraction
Vector database


WebMap
Web
HTML web pages
Snippet extraction
Embedding
LLM
Vector database


WebMap
Web
HTML web pages
Snippet extraction
Embedding
LLM
Vector database
User task description
Most relevant web pages

Snippet Extraction

MUNI
FI

Average document length in tokens

count	143.000000
mean	**33420.923077**
std	**38743.460861**
min	2906.000000
25%	25382.000000
50%	27977.000000
75%	30079.500000
max	314747.000000

(SentencePiece)

MUNI
FI

Average document length in tokens

- **~33 000 tokens** on average (content management website)

MUNI
FI

Average document length in tokens

- **~33 000 tokens** on average (content management website)
- Websites like **reddit or gitlab** on average more than **100 000 tokens**

MUNI
FI

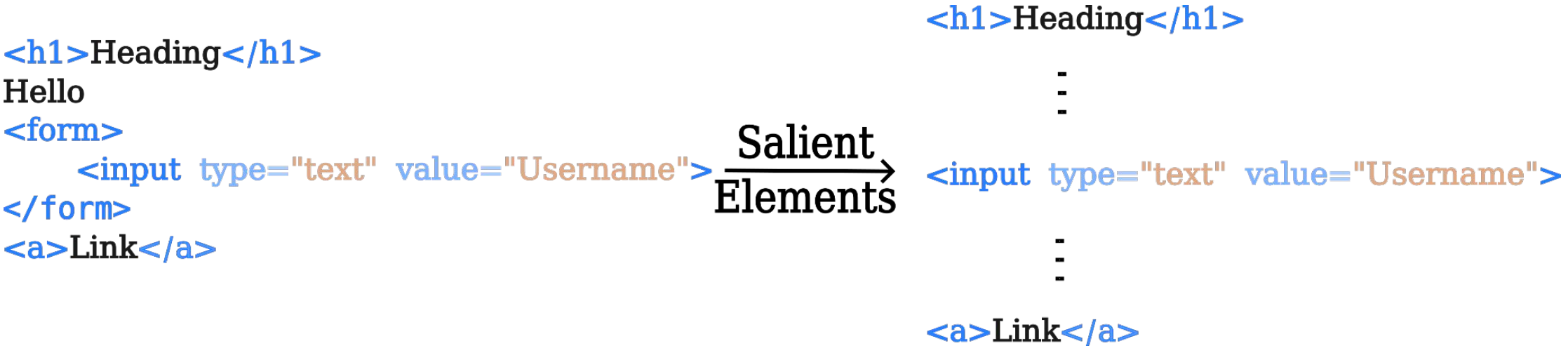
Snippet Extraction

```
<script>...</script>
<h1 class="...">Heading</h1>
<div>
    <span id="...">Hello</span>
    <div>
        <input type="text" value="Username">
        <br>
        <a href="...">Link</a>
    </div>
</div>
```


```
<h1>Heading</h1>
Hello
<input type="text" value="Username">
<a>Link</a>
```

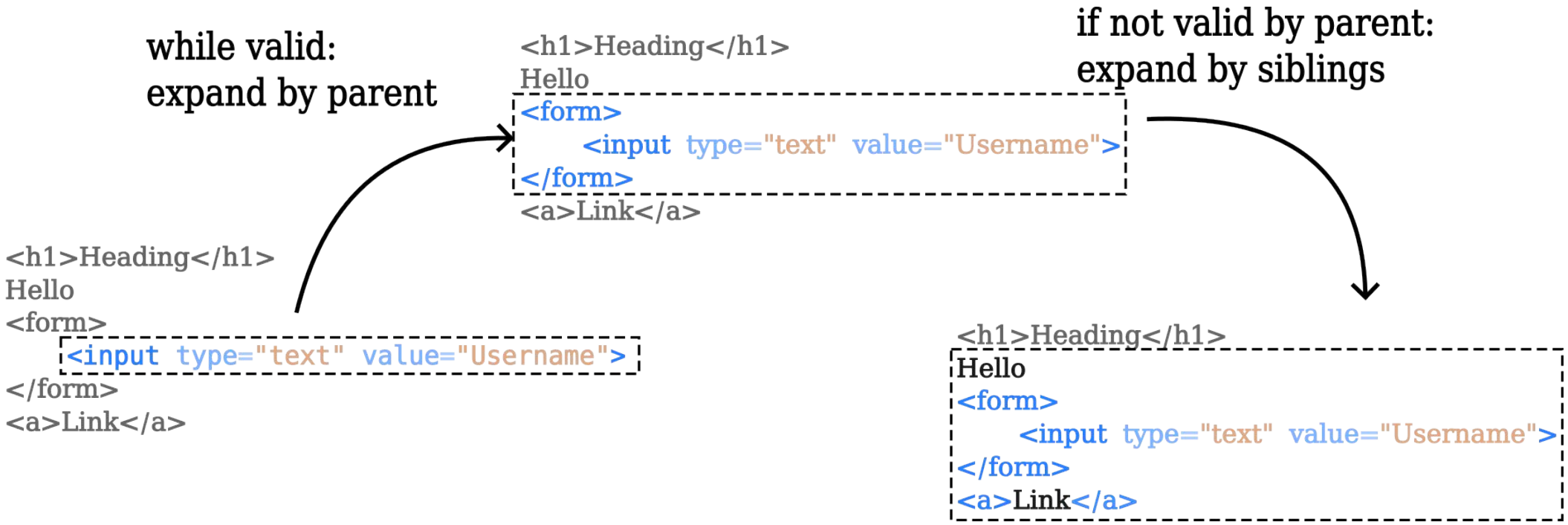
MUNI
FI

Snippet Extraction



MUNI
FI

Snippet Extraction



MUNI
FI

Snippet Extraction Parameters

- Maximum length in characters

MUNI
FI

Snippet Extraction Parameters

- Maximum length in characters
- Maximum Height

MUNI
FI

Snippet Extraction Parameters

- Maximum length in characters
- Maximum Height
- Experimental optimization and evaluation

MUNI
FI

Embedding

- Multiple LLM-based approaches
  - Candidates chosen from the Massive Text Embedding Benchmark

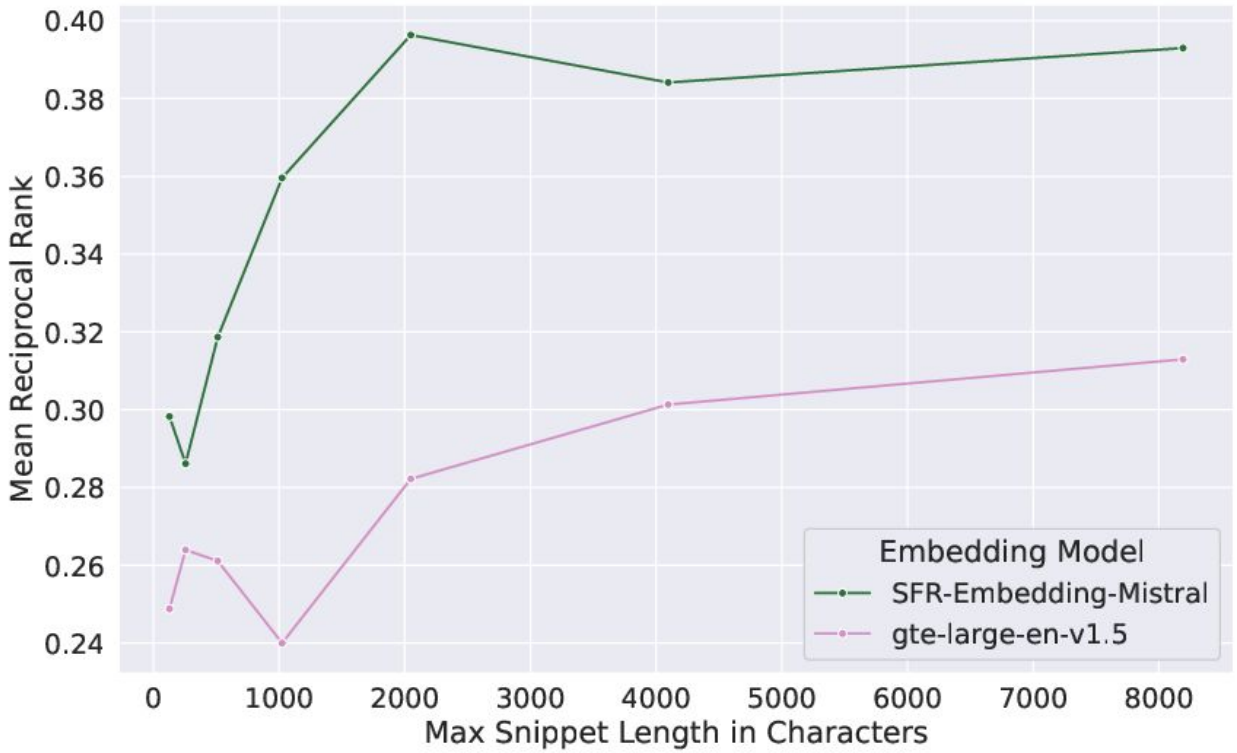
MUNI
FI

Embedding

- Multiple LLM-based approaches
  - Candidates chosen from the Massive Text Embedding Benchmark
- Evaluated using manual human annotations

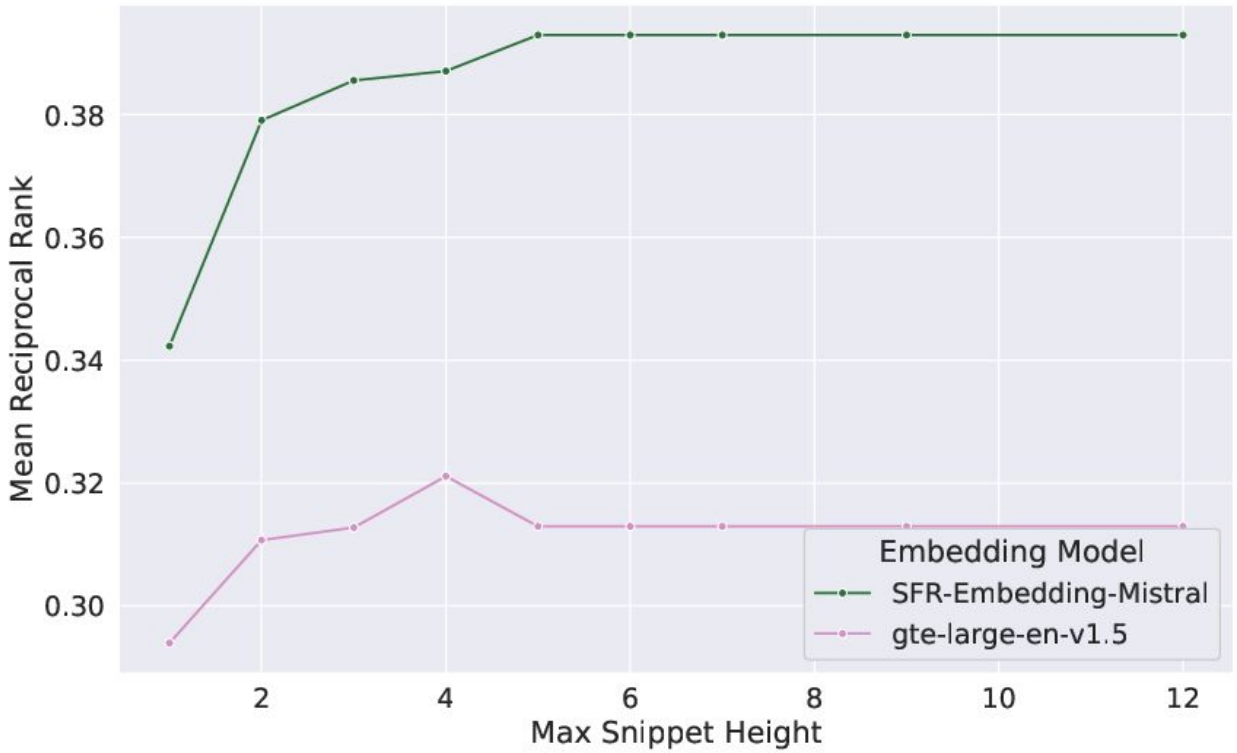
MUNI
FI

Optimal length of a snippet is 2000 characters



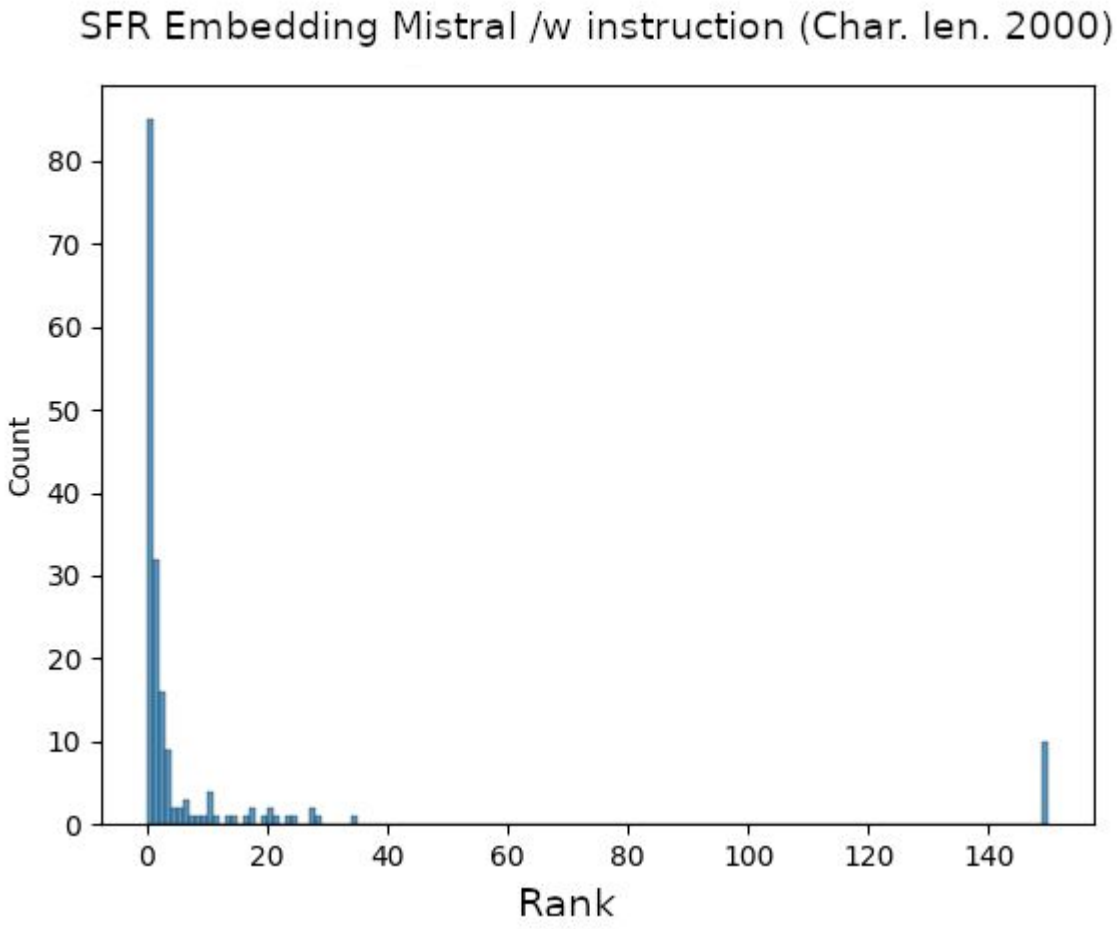
MUNI
FI

Snippet Height is not Significant Parameter



MUNI
FI

Instruction Embedding Performed Best



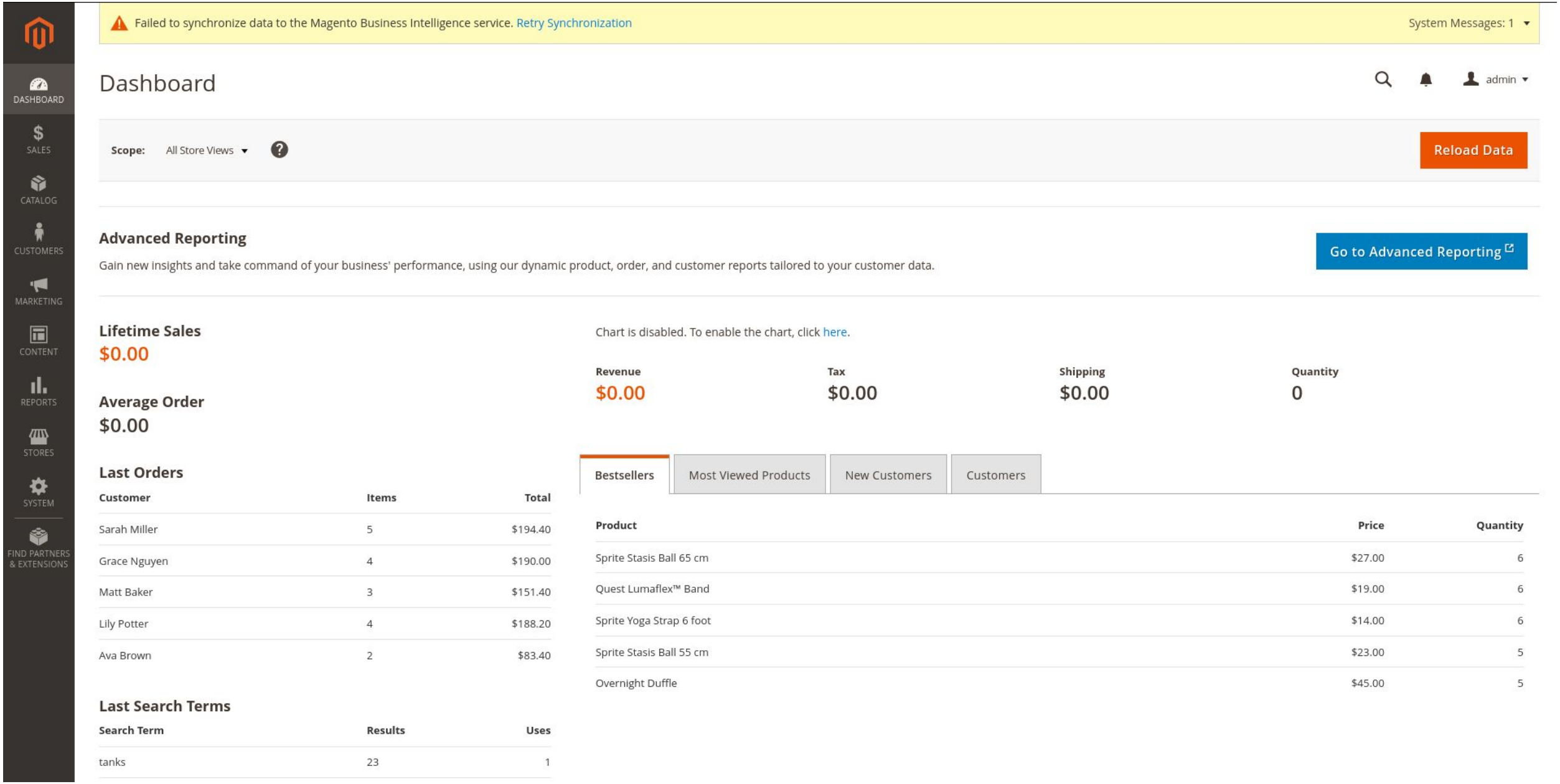
0.61 MRR

MUNI FI

Experiments

- WebArena Evaluation Environment and Dataset
  - ~800 tasks on 5 websites

MUNI
FI


Failed to synchronize data to the Magento Business Intelligence service. Retry Synchronization
System Messages: 1
DASHBOARD
SALES
CATALOG
CUSTOMERS
MARKETING
CONTENT
REPORTS
STORES
SYSTEM
FIND PARTNERS & EXTENSIONS
Dashboard
admin
Scope: All Store Views
Reload Data
Advanced Reporting
Gain new insights and take command of your business' performance, using our dynamic product, order, and customer reports tailored to your customer data.
Go to Advanced Reporting
Lifetime Sales
$0.00
Average Order
$0.00
Chart is disabled. To enable the chart, click here.
Revenue
$0.00
Tax
$0.00
Shipping
$0.00
Quantity
0
Last Orders
Customer Items Total
Sarah Miller 5 $194.40
Grace Nguyen 4 $190.00
Matt Baker 3 $151.40
Lily Potter 4 $188.20
Ava Brown 2 $83.40
Bestsellers
Most Viewed Products
New Customers
Customers
Product Price Quantity
Sprite Stasis Ball 65 cm $27.00 6
Quest Lumaflex™ Band $19.00 6
Sprite Yoga Strap 6 foot $14.00 6
Sprite Stasis Ball 55 cm $23.00 5
Overnight Duffle $45.00 5
Last Search Terms
Search Term Results Uses
tanks 23 1

Baseline

- AutoWebGLM with Gemini 1.5 Pro
  - In-context learning + Chain-of-Thought
  - Simplified HTML of the current web page

MUNI
FI

Baseline

- AutoWebGLM with Gemini 1.5 Pro
  - In-context learning + Chain-of-Thought
  - Simplified HTML of the current web page
- Smaller models usually achieved zero performance
  - Even multimodal model

MUNI
FI

Experimental Results

Method	Accuracy	Relative Gain
AutoWebGLM + Gemini 1.5 Pro (Baseline)	17%	-
AutoWebGLM+Gemini 1.5 Pro+WebMap (aided with human annot.)	23%	30%
AutoWebGLM+Gemini 1.5 Pro+WebMap	20%	15%

*Limited evaluation on a subset of tasks from 1 website

MUNI
FI

Error Analysis

Type of Error	Baseline		Baseline+WebMap (aided with human annot.)	
	Count	Percentage	Count	Percentage
Controls Understanding	24	30	31	40
Incorrect Web Page	18	23	0	0
Observation Understanding	16	20	10	13
Reasoning Error	10	13	6	8
Planning Error	5	6	7	9
Hallucinations	4	5	2	3
Evaluation Dataset Fault	1	1	3	4
Not enough information in the observation	1	1	1	1
Fail to recover from an error	1	1	7	9

Conclusion

- Web navigation is a hard problem

MUNI
FI

Conclusion

- Web navigation is a hard problem
- WebMap **improves** over SOTA by **15%** (in relative gain)
  - Evaluation suggests it could be further improved up to **30%**

MUNI
FI

Conclusion

- Web navigation is a hard problem
- WebMap **improves** over SOTA by **15%** (in relative gain)
  - Evaluation suggests it could be further improved up to **30%**
- Evaluation of **multiple** text **embedding techniques**

MUNI
FI

Conclusion

- Web navigation is a hard problem
- WebMap **improves** over SOTA by **15%** (in relative gain)
  - Evaluation suggests it could be further improved up to **30%**
- Evaluation of **multiple** text **embedding techniques**
- Designing and optimizing a novel **snippet extraction** technique
  - to deal with long web pages

MUNI
FI

Conclusion

- Web navigation is a hard problem
- WebMap **improves** over SOTA by **15%** (in relative gain)
  - Evaluation suggests it could be further improved up to **30%**
- Evaluation of **multiple** text **embedding techniques**
- Designing and optimizing a novel **snippet extraction** technique
  - to deal with long web pages
- **Error analysis highlights** the importance of **modeling in-domain information**

MUNI
FI

Conclusion

- Web navigation is a hard problem
- WebMap **improves** over SOTA by **15%** (in relative gain)
  - Evaluation suggests it could be further improved up to **30%**
- Evaluation of **multiple** text **embedding techniques**
- Designing and optimizing a novel **snippet extraction** technique
  - to deal with long web pages
- **Error analysis highlights** the importance of **modeling in-domain information**
- The thesis will form a basis for a publication in a conference or journal

MUNI
FI

Release Date	Open?	Model Size (billion)	Model	Success Rate (%)
08/2024	✗	Unknown	Jace.AI	57.1
12/2024	✓	-	ScribeAgent + GPT-4o	53
10/2024	✓	-	AgentOccam-Judge	45.7
08/2024	✗	-	WebPilot	37.2
10/2024	✓	-	GUI-API Hybrid Agent	35.8
09/2024	✓	-	Agent Workflow Memory	35.5
04/2024	✓	-	SteP	33.5
04/2024	✓	-	BrowserGym + GPT-4	23.5
04/2024	✓	-	GPT-4 + Auto Eval	20.2
06/2024	✓	-	GPT-4o + Tree Search	19.2
04/2024	✓	7	AutoWebGLM	18.2
06/2023	✓	-	gpt-4-0613	14.9
05/2024	✓	-	gpt-4o-2024-05-13	13.1
06/2023	✓	-	gpt-4-0613	11.7
05/2024	✓	72b	Patel et al + 2024	9.36
03/2023	✓	-	gpt-3.5-turbo-16k-0613	8.87
09/2023	✓	72b	Qwen-1.5-chat-72b	7.14
12/2023	✓	-	Gemini Pro	7.12
04/2024	✓	70	Llama3-chat-70b	7.02
10/2024	✓	7	Synatra-CodeLLama7b	6.28
10/2023	✓	70	Lemur-chat-70b	5.3
03/2024	✓	7	Agent Flan	4.68
08/2023	✓	34	CodeLlama-instruct-34b	4.06
10/2023	✓	70	AgentLM-70b	3.81
04/2024	✓	8	Llama3-chat-8b	3.32
02/2024	✓	7	CodeAct Agent	2.3
10/2023	✓	13	AgentLM-13b	1.6
01/2024	✓	8x7	Mixtral	1.39
10/2023	✓	7	AgentLM-7b	0.74
10/2023	✓	7	FireAct	0.25
08/2023	✓	7	CodeLlama-instruct-7b	0

MUNI FI

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```

```
{
 "sites": [
   "shopping_admin"
 ],
 "task_id": 0,
 "start_url": "http://localhost:7780/admin",
 "intent": "What is the top-1 best-selling product in 2022",
 "eval": {
   "eval_types": [
     "string_match"
   ],
   "reference_answers": {
     "exact_match": "Quest Lumaflex\u2122 Band"
   },
 },
 "most_relevant_link": "/admin/reports/report_sales/bestsellers/"
}
```