

A Multi-Agent Query Fan-Out Analysis of Popular Google Searches in Brazil

Vagner Figueredo de Santana
NIC.br
São Paulo, Brazil
vsantana@nic.br

Beatriz Rocha
NIC.br
São Paulo, Brazil
biarocha@nic.br

Diogo Cortiz
NIC.br / PUC-SP
São Paulo, Brazil
diogo@nic.br

Henrique S. Xavier
NIC.br
São Paulo, Brazil
hxavier@nic.br

ABSTRACT

Large Language Model (LLM) solutions are being more and more embedded into Web search engines, being query fan-out one of them. While decomposing a query and using multiple sub-queries to one or more search engines can increase coverage and diversity, it is not clear how this impact biases in terms of language and location of these sources. In this context, this work employs a multi-agent system to perform a query fan-out study contrasting search engines vs. query fan-out using 10 different search engines via a SearXNG tool provided to agents. Results point out that larger models (e.g., deepseek-r1) tend to answer based on internal knowledge while smaller models invoke tools from 2 up to 4.4 times on average per task. In terms of URLs used by agents to answer queries, even tasks considering Brazilian context and terms in Portuguese, URLs from .com top level domains are used 6.1 times more than .com.br and sources in English are selected over 3.7 times more than non-English sources. With the presented results, we propose the following bias mitigation strategies in agentic Web search: (1) controlling for dominant languages and sources, (2) employing context-aware search engines to better balance priority given to English content, (3) applying prompt engineering techniques to increase transparency and responsibility, and (4) testing different models+frameworks for revealing differences in terms of token usage, query augmentation, and tool usage for different solutions.

KEYWORDS

Artificial Intelligence, Agentic Web, Large Language Models, LLM-based Agents, Multi-Agent Systems, Query Analytics

1 INTRODUCTION

Understanding how information sources are selected, ranked, and aggregated for a search query directly impacts the quality, diversity, and framing of the information consumed by

users interacting with Web search engines. Ranking and retrieval pipelines influence exposure to information, visibility of sources, and downstream user behavior [3]. In addition, search systems may reproduce societal biases, amplify dominant perspectives, and affect user trust and decision making [10, 21]. Consequently, understanding how these search systems collect and mediate information has become central to assessing transparency, fairness, and information quality on the Web.

More recently, search engines have been transitioning from traditional information retrieval (IR) pipelines towards LLM-based search experiences. Instead of only retrieving ranked hyperlinks, modern systems increasingly synthesize information through generative interfaces powered by LLM-based agentic workflows. This transition changes the interaction paradigm from document retrieval to summarization and answer generation, in which retrieved sources are processed and reformulated before being presented to users. This shift introduces new concerns regarding hallucination, factuality, provenance, transparency, and the amplification of latent biases inherited from training corpora [16]. Concurrently, search platforms have adopted increasingly sophisticated query expansion and decomposition strategies, referred to as *query fan-out* [4]. While this approach aims at improving retrieval coverage, it alters the sourcing mechanisms of search engines, negatively impacting transparency and provenance.

Query fan-out refers to the process of decomposing, reformulating, or augmenting an input query into multiple derived queries that are independently executed against one or more retrieval systems [4, 12]. In LLM-based search systems, these operations are often dynamically generated during inference time, closely connecting modern search with the notion of AI (Artificial Intelligence) agents.

In classical AI literature, an agent is “*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators*” [17]. With recent advances in AI and LLMs, the goals envisioned with the idea of *perceiving* and *acting* were augmented by a natural language space [27]. The capabilities brought by LLMs applied to AI agents powered the now popular LLM-based agents. These agents may act on model of their environments,

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia’2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

learn about their operation, or involve complex orchestration [24], being more useful for tasks that are harder, more realistic, and usually don't have a single correct answer [9]. LLM-based systems increasingly employ LLM-based agent workflows capable of decomposing user tasks, coordinating multiple retrieval actions, interacting with tools, and aggregating intermediate results into synthesized outputs.

In modern systems, autonomous agents do not operate alone but as **multi-agent systems** (MAS), coordinating, negotiating, or delegating tasks among a team of specialized agents. Some agents act as "*orchestrators*" that manage workflows, assign sub-tasks, or monitor progress across sub-agents. These agents combine perception, planning, memory, and coordination faculties to drive large-scale or hierarchical task completion [24]. MAS have also proven value for open-ended research tasks and for breadth-first queries, pursuing multiple directions simultaneously [1].

In this context, this research aims to investigate how LLM-powered search engines impact data quality and biases considering Google popular searches performed in Brazil. In particular, this work studies the following research question (RQ):

RQ1 How do query fan-out searches differ from those performed directly via search engines in terms of language biases and sources of information?

1.1 Contribution Statement

This work contributes with a query fan-out study employing a factorial experimental design combining models $\times$ agentic frameworks $\times$ tasks, i.e., ($4 \times 2 \times 20 = 160$ task executions), as follows:

- Four open-weight large language models (two baseline models and two state-of-the-art models),
- Two agentic implementations using different frameworks (CrewAI [5] and LangGraph [11]),
- 20 tasks derived from the most common searches performed on Google in Brazil in 2025 [8].

The study evaluates how different agentic search configurations influence retrieval behavior, generated responses, and the propagation of biases across multilingual and multi-step search interactions.

2 RELATED WORK

Bias in search involve activity bias, reflected by different participation in the Web, and algorithmic bias added during the design and implementation of systems [2]. Moreover, search engines do not cover diverse viewpoints and suffer from ideological bias, favoring an individual ideological view to others, which impacts towards polarization of our society [13]. In a study on confirmation bias in Web search, literacy on a topic has shown to be a protective measure on supporting users to examine search results and explore different opinions [25]. In terms of gender bias for job searches, a study has shown that search engines count on different page ranks and, sometimes, skew towards "male-dominant" positions for both men and women [26]. Literature also presents existing

bias in LLM-based systems and how undersampling is a potential solution to deal with this issue [6]. Moreover, prompt engineering has also shown to be a tool to address fairness issues [14, 20].

Generative AI (GenAI) does not search, it selects information [15]. In a study focused on how information is selected by ChatGPT in the Brazilian context, the following was found [15]:

- (1) **Lack of diversity:** few domains are considered in the answer;
- (2) **URL findability:** URLs that contain keywords are better ranked;
- (3) **Lifestyle bias:** professional and technical content have priority over entertainment;
- (4) **Language bias:** even in business tasks defined in Portuguese, English content is in the majority of the responses;
- (5) **Traditional Web as GenAI infrastructure:** Wikipedia is still an important source for factual content.

Literature suggests that GenAI search operates in a two-layer fashion [12]: (1) A layer that decides whether to search (i.e., using internal knowledge vs. searching for information); (2) A query-decomposition layer that determines what to search for.

This creates a new indirection layer for users interacting with Web searches and, for MAS, this can be yet another layer reducing transparency, provenance, and potentially amplifying biases. In this context, we study how MAS powered by different models influence responses and references considered in the responses when considering a specific scenario of common searches in Brazil.

In terms of MAS architecture, a study proposes an architecture for moderation of textual responses composed of agents operating in an iterative review workflow with shared context, including semantic and contextual evaluation, improvement recommendations, textual rewriting, and final decision-making [7].

Related work on assessing query fan-out shows that user intent influences fan-out composition, discovery queries trigger 3.3 times the entity injection rate of informational queries, platforms exhibit distinct retrieval behaviors in terms of injecting entities on queries, user intent and platform are the dominant factors for query fan-out, and that search trigger rate varies, suggesting larger models are more confident in answering from training data alone [12].

3 METHOD

This study involves performing query fan-out searches using MAS powered by four open weight models and implemented using two different agentic frameworks (CrewAI [5] and LangGraph[11]). The terms considered in the queries are the top 20 search terms on Google originating from Brazilian users in 2025 [8].

The selection of models involved open weight models and baseline models, with higher probability of tool use. The limit for selecting these models also considered the capabilities of

the in-house server in which Ollama was run. The referred server is powered by an NVIDIA RTX 6000 Ada Generation (48G single GPU). Hence, the following Ollama's model ids were used in the experiment:

- *deepseek - r1 : latest* (8B)
- *gemma4 : latest* (8B)
- *qwen2.5 - 7B : latest*
- *granite4.1 : 3b*

In order to test different GenAI models and agentic frameworks, the testing framework utilized was MASTIF (Multi-Agent System TestIng Framework) [18]. MASTIF is an open source MAS testing framework developed by Web Technologies Study Center (Ceweb.br), Brazilian Network Information Center (NIC.br), and it is available on GitHub¹. The goal of MASTIF is to support the specification of multiple tasks, a prompt template, connection with different inference providers (e.g., HuggingFace and OpenAI, Anthropic, Ollama), and testing of multiple models and implementation frameworks. The testing framework logs latency, token, intermediate reasoning steps, and other metrics to support empirical studies of MAS on the Web.

The web search tool provided to agents was developed using SearXNG API [23]. The SearXNG server was hosted locally and the following search engines configured: *google*, *bing*, *duckduckgo*, *brave*, *startpage*, *quant*, *mojeek*, *wikipedia*, *reddit*, and *github*. SearXNG was used under default configuration setting for language, i.e., automatic detection.

The workflow configured in the testing framework involved 3 agents:

- (1) **Web Search**, responsible for performing the baseline searches.
- (2) **Metasearch**, responsible for using SearXNG API for sub-querying and augmenting searches, and
- (3) **Orchestrator**, responsible for compiling results retrieved by previous agents.

Next, we detail the prompt for these 3 agents in the considered workflow. Placeholders are populated by the testing framework and results from each agent are concatenated to `{research_results}` placeholder.

Web Search

Role: {role}

Framework: {framework}

TASK: {task}

Your objective is to establish a search baseline.

Use the 'web_search' tool to find the most prominent and direct information regarding the query.

DO NOT SIMULATE RESULTS.

DO NOT PROVIDE HYPOTHETICAL RESULTS.

Instructions:

1. Use the available tools: {tools_text}
2. Focus on the top 3-5 high-authority results.
3. Note the primary intent of the search (e.g., navigational, informational, transactional).

¹<https://github.com/cewebbr/mastif>

4. Provide a concise summary of your findings.

5. Output a list of URLs found to 'research_results' under the identifier key *web_search*.

Your output will be appended to the cumulative 'research_results' list for later analysis.

Metasearch

Role: {role}

Framework: {framework}

TASK: {task}

CURRENT STATE OF RESEARCH: {research_results}

Your objective is to broaden the research using the 'searxng_search' tool. This tool aggregates results from multiple engines (Google, Bing, DuckDuckGo, etc.).

DO NOT SIMULATE RESULTS.

DO NOT PROVIDE HYPOTHETICAL RESULTS.

Instructions:

1. Execute a broad search using 'searxng_search'.
 2. Identify any information or perspectives that were missing from the baseline research provided above.
 3. Crucially, look for "Engine Attribution" in the tool output. Note which engines provided which results.
 4. If the results are identical to the baseline, state that the search indices are overlapping.
 5. Output a list of URLs found to 'research_results' under the identifier key *metasearch*.
- Expand the 'research_results' with these diversified findings.

Orchestrator

Role: {role}

Framework: {framework}

Step: {step}/{max_steps}

FINAL SYNTHESIS TASK: {task}

Below is the complete log of findings gathered from the baseline search and the metasearch fan-out:

{research_results}

Your goal is to provide the final experiment assessment.

Please address:

1. **COMPREHENSIVE ANSWER:** Provide the best possible answer to the original task based on all gathered data.
 2. **TRAFFIC ANALYSIS:** Compare the volume and variety of information retrieved by the standard search vs. the metasearch fan-out.
 3. **ENGINE QUALITY:** Based on the attribution, which search engines (e.g., Google, Bing, Wikipedia) provided the most relevant "niche" results?
 4. **OVERLAP ASSESSMENT:** Did the fan-out significantly increase the "breadth" of information, or was it redundant?
 5. **REFERENCES:** List all URLs found in *web_search* and *metasearch* properly identified.
- Format your response as a professional research report.
-

MASTIF also supports the standardization of empirical studies by defining multiple parameters in a YAML file including *temperature* for model variability, *max_tokens* per response, *max_steps* per agent,

max_tool_rounds for limiting tool use by agents, and response *timeout*. The following parameters were considered in this experiment:

- *temperature*: 0.7, aiming at diverse search interpretation; lower temperature tends to fix behavior on how agents use tools and higher temperature tends to allow different behaviors in terms of how to perform a task, tool use, including tool call hallucinations.
- *max_tokens*: 2048, to support cumulative {research_results}.
- *max_steps*: 5, to allow multiple agent loops if needed.
- *max_tool_rounds*: 5, to allow agents to use one or more tools multiple times.
- *response*: 60, increased timeout to handle automation restrictions placed by search engines.

Next, we detail the prompt used in the {*task*} placeholder used in the previous agent prompts. The {*query*} placeholder was used to host the search terms from the top searches on Google originating from Brazil, in 2025. Prompts were created following good practices and recommendations from [19, 22].

Task

You are a Search Quality Analyst. Your goal is to execute the following search task:

Identify the primary intent and top information sources for: {query}

Analyze the results for accuracy, relevance to the Brazilian context, and engine-level performance.

Provide your analysis based on the cumulative findings stored in the workflow, including information sources, language of the content, timestamps, and any relevant metrics.

The 20 top searched terms from [8] were the following (alphabetical order): *Canais Globo*, ChatGPT, *Clima*, Crefisa, Enem, Facebook, G1, Gmail, Google *Tradutor*, Hotmail, Instagram, *Loterias*, *Mercado Livre*, *Previsão do tempo*, Roblox, Shopee, *Tabela FIPE*, UOL, WhatsApp Web, and YouTube.

Finally, in terms of data analysis, the following metrics were considered: reasoning tokens, output tokens, tool calls, average duration of tool calls, query augmentation by agents, top level domains considered in the answers, content languages, and quality of the reports generated. For reproducibility, the YAML file used in this experiment can be found at MASTIF's repository ².

4 RESULTS

4.1 Token Usage

Given the testing framework configuration, all workflows had a similar number of steps and reasoning token usage. Table 1 shows that differences were more consistent in terms of framework implementation, with LangGraph using an average

of 23.25 less tokens in reasoning steps than CrewAI. Kruskal-Wallis test shows that reasoning tokens differ across models (H-statistic = 58.338, p-value < 0.001) and that they do differ in terms of frameworks (H-statistic = 71.915, p-value < 0.001). Dunn's post-hoc test points that the following models present a significant pairwise difference: deepseek-r1 and gemma4 (p < 0.001), deepseek-r1 and granite4.1 (p < 0.001), and deepseek-r1 and qwen2.5-7B (p < 0.001).

Significant differences were observed also in terms of output tokens. Table 2 details the mean number of tokens used by each model+framework per task. Qwen2.5-7B was the model that used less tokens (around 1k tokens) on average, while other models used from 1.2k tokens up to 1.6k tokens on average per task. No significant difference was observed when comparing CrewAI and LangGraph in terms of token usage. Kruskal-Wallis test shows that reasoning tokens differ across models (H-statistic = 48.626, p-value < 0.001) and that they do not differ in terms of frameworks (H-statistic = 1.291, p-value = 0.256). Dunn's post-hoc test points that the following models present a significant pairwise difference: deepseek-r1 and qwen2.5-7B (p-value < 0.001), gemma4 and qwen2.5-7B (p-value < 0.001), and granite4.1 and qwen2.5-7B (p-value < 0.001).

Table 1: Mean reasoning tokens per task, model+framework.

Model+framework	Mean (token)	σ (token)
deepseek-r1+CrewAI	267.90	3.65
deepseek-r1+LangGraph	243.90	3.65
gemma4+CrewAI	241.90	3.24
gemma4+LangGraph	218.90	3.24
granite4.1+CrewAI	241.90	3.24
granite4.1+LangGraph	218.90	3.24
qwen2.5-7B+CrewAI	241.75	3.06
qwen2.5-7B+LangGraph	218.75	3.06

Table 2: Mean token usage for generating outputs per task, model+framework.

Model+framework	Mean (token)	σ (token)
deepseek-r1+CrewAI	1610.90	504.34
deepseek-r1+LangGraph	1280.95	569.57
gemma4+CrewAI	1430.70	142.01
gemma4+LangGraph	1404.25	187.38
granite4.1+CrewAI	1588.40	523.83
granite4.1+LangGraph	1652.05	318.41
qwen2.5-7B+CrewAI	1025.30	228.70
qwen2.5-7B+LangGraph	1049.05	209.23

4.2 Tool Use

In terms of tool use by agents, Table 3 details the average tool calls per task by model+framework. While deepseek-r1

²<https://github.com/cewebbr/mastif/blob/main/experiments/queryfanout.yaml>

had no calls and models often triggered around two tool requests, qwen2.5-7B+CrewAI triggered 3.15 and qwen2.5-7B+LangGraph triggered 4.45 requests on average. Kruskal-Wallis test shows that tool calls differ across models (H-statistic = 131.377, p-value < 0.001) and that they do not differ in terms of frameworks (H-statistic = 1.686, p-value = 0.194). Dunn’s post-hoc test points that models present a significant pairwise difference: deepseek-r1 and gemma4 (p-value < 0.001), deepseek-r1 and granite4.1 (p-value < 0.001), deepseek-r1 and qwen2.5-7B (p-value < 0.001), gemma4 and qwen2.5-7B (p-value < 0.001), and granite4.1 and qwen2.5-7B (p-value < 0.001).

Table 3: Average tool calls per task by model+framework

Model+framework	Mean (calls)	σ (calls)
gemma4+CrewAI	2.00	0.00
gemma4+LangGraph	2.05	0.22
granite4.1+CrewAI	2.10	0.72
granite4.1+LangGraph	2.35	0.49
qwen2.5-7B+CrewAI	3.15	0.88
qwen2.5-7B+LangGraph	4.45	1.10

Regarding average time taken in tool requests in our setup, Table 4 shows that gemma4 took almost twice of the time per request than smaller models such as granite4.1 and qwen2.5-7B. Kruskal-Wallis test shows that time running tools differs across models (H-statistic = 129.153, p-value < 0.001) and that it does not differ in terms of frameworks (H-statistic = 0.566, p-value = 0.452). Dunn’s post-hoc tests point that models present a significant pairwise difference: deepseek-r1 vs gemma4 (p-value < 0.001), deepseek-r1 vs granite4.1 (p-value < 0.001), deepseek-r1 vs qwen2.5-7B (p-value < 0.001), gemma4 vs granite4.1 (p-value < 0.001), and gemma4 vs qwen2.5-7B (p-value < 0.001).

Table 4: Average time taken when running tools per task, by model+framework

Model+framework	Mean (ms)	σ (ms)
gemma4+CrewAI	900.56	105.16
gemma4+LangGraph	858.50	83.00
granite4.1+CrewAI	561.26	136.13
granite4.1+LangGraph	590.49	186.17
qwen2.5-7B+CrewAI	629.42	132.04
qwen2.5-7B+LangGraph	502.84	102.46

4.3 Query Augmentation

Table 5 shows model+framework setups that augmented input queries, i.e., agents tried to add terms to gather more precise, contextualized answers. In this regard, granite4.1 (both frameworks) was the most consistent in augmenting queries. It added context on around one third of calls across both frameworks. Augmentation strategies included:

- (1) Geographic disambiguation: “Clima Brazilian weather app”, “Loterias Brazil”, “Previsão do tempo Brazil”;
- (2) Semantic disambiguation: “G1 Brazilian context”, “G1 meaning”, “UOL Universo Online Brazilian company”, “UOL Brazilian online payment platform”.

Figure 1 details all 41 augmentation instances by model+framework. It is possible to verify cases in which models hallucinated, for instance, granite4.1+CrewAI augmented query for ‘UOL’ (a news portal) to ‘Unibanco Holding S.A.’ (a bank) and ‘G1’ (another news portal) to ‘G1 CNC G-code Brazilian context’ (CNC programming language appending ‘Brazilian context’). Interesting patterns also were identified towards certain terms, for instance, granite4.1 and qwen2.5-7B augmented ‘Facebook’ queries with technical issues terms, for instance, ‘Facebook technical issues’, ‘Facebook Help Center Brazilian’, and ‘Facebook error performing query troubleshooting Brazilian.’ Other than these cases, models and frameworks augmented queries aiming at retrieving content for Brazilian context, as asked in the prompt provided to agents.

Qwen2.5-7B+LangGraph — augmented only in LangGraph, representing a framework effect, i.e., the same model issued 0 augmented queries under CrewAI but 15 under LangGraph, revealing the value of testing in more than one framework implementation. Augmentations in qwen2.5-7B+LangGraph were simpler, mostly appending “Brazil” to search terms (e.g., “Clima Brazil”, “Roblox Brazil”, “Shopee Brazil”) or restating intent (“Mercado Livre Brazil primary intent”, “Facebook technical issues”).

Deepseek-r1 did not triggered tool requests and gemma4 had zero augmentation in both frameworks. Despite producing the best analytical output, gemma4 never modified the query. It always searched the exact task term and derived its analysis from the results returned.

Table 5: Query augmentation per model+framework.

Model+framework	Searches	Aug.	Rate(%)
granite4.1+CrewAI	39	15	38.5%
granite4.1+LangGraph	40	11	27.5%
qwen2.5-7B+LangGraph	74	15	21.7%
Total	153	41	26.8%

4.4 Sources of Information

When assessing the use of .com sources by model and framework, Kruskal-Wallis test shows that the use of information from .com sources differ across models (H-statistic = 46.833, p-value < 0.001) and that they do not differ in terms of frameworks (H-statistic = 0.074, p-value = 0.785). Dunn’s post-hoc tests points that the following models present a significant pairwise difference: deepseek-r1 and gemma4 (p-value < 0.001), deepseek-r1 and qwen2.5-7B (p-value < 0.001), and gemma4 and granite4.1 (p-value < 0.001).

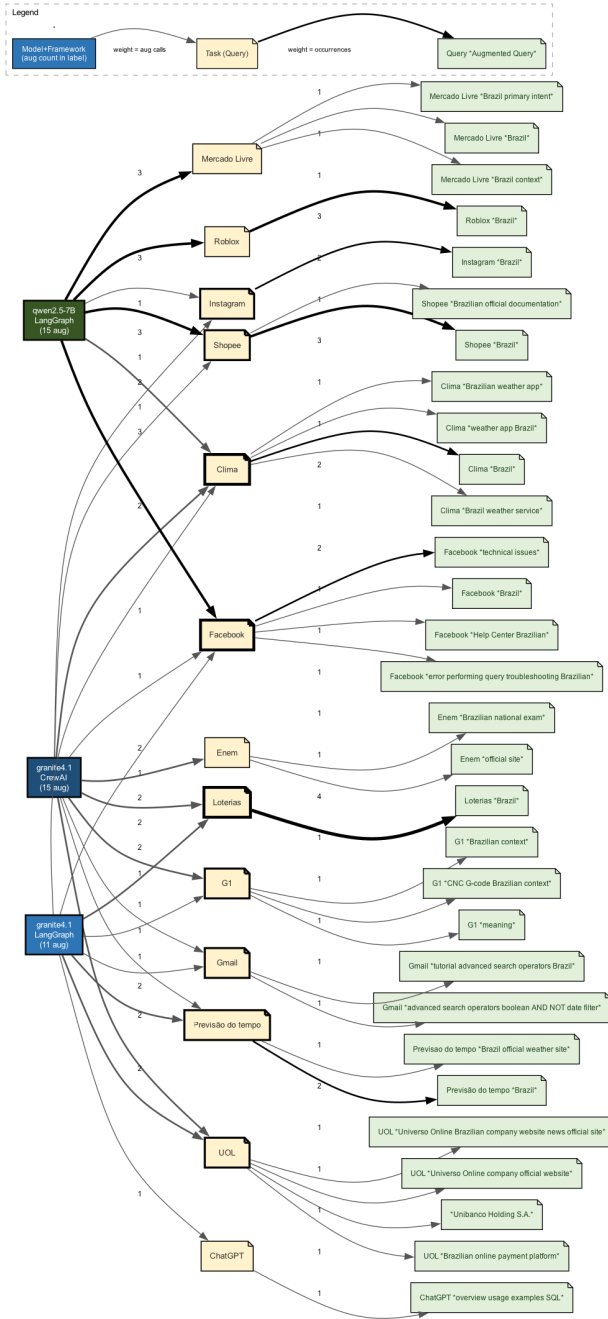


Figure 1: Tripartite graph depicting model+framework (left), terms with query augmentations (center), and the augmented queries (right).

When assessing the use of *.com.br* sources by model and framework, Kruskal-Wallis test shows that the use of information from *.com.br* sources do not differ across models (H-statistic = 2.625, p-value = 0.453) and that they do not differ in terms of frameworks (H-statistic = 0.122, p-value = 0.727).

In terms of URLs used by agents to answer queries, even tasks considering Brazilian context and terms in Portuguese, URLs from *.com* top level domains are used in answers 6.1 times more used than *.com.br*, on average (Figure 2). Table 6 details the model+framework reliance on content from *.com* websites vs. *.com.br*. Gemma4 and qwen2.5-7B were the models with strongest bias towards *.com* websites vs. *.com.br*. Qwen2.5-7B+CrewAI relied on *.com* websites over 15 times more than *.com.br* ones.

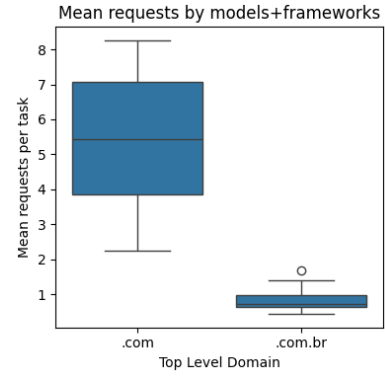


Figure 2: Mean requests by model+framework by top level domain.

In addition, when assessing content language, sources in English are selected over 3.7 times more than non-English sources on average (Figure 3). Table 7 contrasts the use of content in English with non-English content. With only one exception being deepseek-r1+CrewAI, which relied more on non-English content (Portuguese plus Spanish), all models+framework combinations relied more on English content to perform the tasks. Qwen2.5-7B was the model that relied more on English content, considering up to 10 times more English content than non-English content while performing the tasks (qwen2.5-7B+LangGraph).

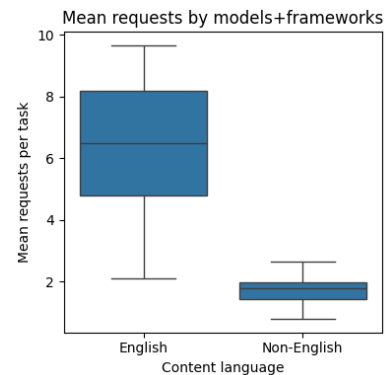


Figure 3: Mean requests by model+framework by content language.

Table 6: Average number of top level domains used in answers ending in .com vs .com.br, by task, model+framework.

Model+framework	Mean .com (σ)	Mean .com.br (σ)	Fold
deepseek-r1+CrewAI	2.25 (2.90)	1.40 (2.44)	1.61
deepseek-r1+LangGraph	3.10 (3.48)	1.70 (2.41)	1.82
gemma4+CrewAI	8.25 (2.67)	0.75 (1.29)	11.00
gemma4+LangGraph	7.70 (2.92)	0.85 (1.50)	9.06
granite4.1+CrewAI	4.60 (3.39)	0.70 (1.26)	6.57
granite4.1+LangGraph	4.10 (2.65)	0.70 (1.26)	5.86
qwen2.5-7B+CrewAI	7.00 (4.04)	0.45 (0.69)	15.56
qwen2.5-7B+LangGraph	6.70 (3.84)	0.45 (0.69)	14.89
Mean	4.99 (3.20)	0.81 (1.40)	

Table 7: Content language used for performing tasks, by model+framework.

Model+framework	Mean English (σ)	Mean non-English (σ)	Fold
deepseek-r1+CrewAI	2.10 (3.04)	2.25 (3.74)	0.93
deepseek-r1+LangGraph	3.45 (4.02)	2.65 (3.94)	1.30
gemma4+CrewAI	9.65 (3.51)	1.90 (2.77)	5.08
gemma4+LangGraph	8.70 (3.42)	1.90 (2.71)	4.58
granite4.1+CrewAI	5.25 (3.64)	1.65 (2.21)	3.18
granite4.1+LangGraph	5.30 (3.23)	1.60 (2.33)	3.31
qwen2.5-7B+CrewAI	7.65 (3.94)	1.00 (1.78)	7.65
qwen2.5-7B+LangGraph	8.00 (4.14)	0.80 (1.32)	10.00
Mean	6.21 (3.66)	1.68 (2.53)	

4.5 Results Summary

In sum, when comparing **models**, significant differences were found in terms of (1) reasoning tokens, (2) output tokens, (3) tool usage, and (4) tool running time. When accessing different **implementation frameworks**, significant differences were found only in terms of reasoning tokens. Beyond that, framework was also identified as an effect impacting the **query augmentation**, given different results from qwen2.5-7B+CrewAI and qwen2.5-7B+LangGraph.

5 DISCUSSION

5.1 Query Augmentation

Query augmentation has shown to be model-specific and framework-specific. For instance, tasks on ambiguous terms (e.g., ‘G1’, ‘UOL’, and ‘Clima’) triggered augmentation in granite4.1, but not in gemma4 or qwen2.5+CrewAI. Granite4.1 was the only model to consistently reason that short or ambiguous Brazilian terms needed context before searching. The LangGraph framework appears to mildly encourage augmentation (qwen2.5 augmented only in LangGraph), but this effect was not observed in gemma4, suggesting that this is a two-factor (model and framework) specific.

5.2 Search Grounding vs. Hallucination

Deepseek-r1+CrewAI was the most transparent about failures as it explicitly labeled simulated content while answering

based on its internal knowledge [12]. Granite4.1 (both frameworks) was the most grounded model+framework tested as tools returned no relevant results (e.g., dictionary definitions for ‘G1’, lexical entries for ‘UOL’), it reported exactly that rather than substituting expected content. Gemma4+CrewAI occasionally added analysis that went beyond what the search tool returned (e.g., suggesting INMET (Brazilian National Institute of Meteorology) while searching for ‘Clima’), but distinguished inference from observed data. Qwen2.5-7B (both frameworks) maintained consistent grounding, reporting search tool outputs even when they were irrelevant for the query at hand.

5.3 Localization and Brazilian Context

All models consistently found English-language results dominating search outputs, even for distinctly Brazilian queries like ‘Canais Globo,’ ‘Loterias,’ ‘Tabela FIPE,’ and ‘Enem.’ Gemma4 was the most proactive in diagnosing this as a localization gap and recommending Portuguese-language query strategies. Granite4.1 noted the absence of Brazilian-domain sources in results, e.g., .com.br and .gov.br. No model successfully found INMET for the ‘Clima’ query, which can be seen as a signal that the search results were not well-calibrated for Brazilian-context. ‘G1,’ a major Brazilian news portal, was never identified as the intended target when searching ‘G1.’ All models reported technical disambiguation failures (e.g., G-code CNC language).

5.4 Metasearch Fan-Out Value Assessment

Across the models that produced usable output, the following patterns emerged in terms of the value of performing fan-out queries:

- (1) Fan-out increased conceptual breadth (new intent categories found).
- (2) Fan-out increased URL count but with high overlap at top ranked URLs.
- (3) Fan-out surfaced more Portuguese/regional sources.

5.5 CrewAI vs. LangGraph

CrewAI generally produced longer, richer reports with more elaborate formatting. LangGraph outputs were more uniform in length across models, suggesting that the framework follows output constraints. The explicit Web search/metasearch splits (gemma4, granite4.1) occurred predominantly in CrewAI, suggesting CrewAI's agent orchestration better preserved the two-step tool call structure for this experiment. Reasoning token allocation was consistently lower in LangGraph in comparison with CrewAI, suggesting LangGraph's specific functioning replaces some of the reasoning.

6 CONCLUSIONS

Connecting back to RQ1 (How do query fan-out searches differ from those performed directly via search engines in terms of language biases and sources of information?), results suggest that there are language and Localization biases in agentic Web search. Models tested that produced usable output flagged the same problem independently: despite all 20 tasks being distinctly Brazilian queries, search results were overwhelmingly originating from English content. Websites from *.com.br*, *.gov.br*, and *.org.br* top level domains were far less frequently than *.com* domains. This was true for both baseline and fan-out calls. The fan-out did not fix this completely. When granite4.1 explicitly augmented queries with "Brazil" or "Brazilian context," it sometimes surfaced more Portuguese-language content, as a result of query augmentation. Models that issued bare queries got the same English-heavy results in both passes. This suggests a structural bias: SearXNG configuration and/or the underlying search engines it primarily draws from are not well-calibrated for Brazilian Portuguese informational queries.

Both approaches faced hurdles with disambiguation. Queries such as 'G1,' 'UOL,' 'Tabela FIPE,' and 'Enem' require cultural context to interpret them correctly. This disambiguation failure was identical in the baseline and the fan-out. No fan-out call returned the correct Brazilian entity for 'G1' or 'UOL.'

Fan-out did add value in terms of depth, given that were in which fan-out broadened the information landscape. For navigational queries with rich ecosystems (e.g., 'Facebook,' 'Shopee,' 'Roblox,' and 'Mercado Livre'), the additional calls surfaced developer-oriented content such as GitHub repositories, API documentation, and scraping guides, going beyond the baseline searches. A search engine query with added terms

like 'API,' 'developer,' 'download') would achieve a similar expansion without the MAS overhead. In navigational queries (e.g., 'YouTube,' 'WhatsApp Web,' and 'Gmail'), fan-out produced the higher overlap with the baseline.

Agent layer introduced context through query augmentation. Granite4.1 (both frameworks) and qwen2.5+LangGraph augmented the query with terms like 'Brazil,' 'Brazilian context' produced more localized results. This has potential to improve queries, but, at the same time, to add biases in an opaque way.

Based on presented results, mitigation strategies for tackling localization and language bias in agentic Web search should include:

- (1) Controlling for dominant languages and sources [15], for instance, via undersampling [6];
- (2) Ensure that MAS-powered search employs existing context-aware mechanisms to better balance priority given to English content [15],
- (3) Employing prompt engineering techniques to increase transparency and responsibility [14, 20], and
- (4) Testing models+frameworks for revealing token usage, query augmentation, and tool usage before going to production as content retrieved shall differ in terms reasoning and outputs, query augmentation, and sources of information considered.

Performing empirical studies on the agentic Web is a concrete way to tap into this complex software engineering task of developing agentic workflows that may or may not use available tools, connecting with models that may or may not hallucinate, and all of them dependent on textual content embedded in the base models.

In terms of limitations, this study aimed at contrasting fan-out and web searches using a MAS. The models selected were constrained by the hardware available for running the study. Models selected aimed to cover open weight models of different sizes, allowing for different tool usage capabilities. Moreover, tasks were specific to terms Brazilians searched on Google in 2025, as this study aimed to cover how Brazilians search for information on current LLM-based search engines. The set of query is limited to mostly short, popular terms, restricting generalization of the results to more complex search scenarios. Lastly, in this study, queries (and sub-queries) were performed in an automated way by agents. Hence, the same searches when performed by humans on websites, while providing more profile information, may differ in terms of context and recommendations. However, this work maintains validity in terms of assessing how short queries are performed in a MAS setup in which agents can use tools to perform tasks, connect to different search engines via SearXNG, and augment queries autonomously.

Future steps of this research involve amplifying the experiment for more complex queries, prompt improvements, more powerful models, different localization settings, hyperparameters grid search, and more complex agentic workflows (e.g., hierarchical).

REFERENCES

- [1] Anthropic. 2025. How we built our multi-agent research system. Technical specification. <https://www.anthropic.com/engineering/multi-agent-research-system>
- [2] Ricardo Baeza-Yates. 2020. Bias in Search and Recommender Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) (*RecSys '20*). Association for Computing Machinery, New York, NY, USA, 2. <https://doi.org/10.1145/3383313.3418435>
- [3] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) (*SIGIR '18*). Association for Computing Machinery, New York, NY, USA, 405–414. <https://doi.org/10.1145/3209978.3210063>
- [4] Conductor. 2025. *What Is Query Fan-Out in AI Search and Why Does It Matter?* Technical Report. Conductor. <https://www.conductor.com/academy/query-fan-out/> Accessed: 2026-05-29.
- [5] CrewAI. [n. d.]. CrewAI. <https://github.com/crewAIInc/crewAI> Accessed: 2025-09-19.
- [6] Guilherme Fonseca, Gabriel Prenassi, Washington Cunha, Marcos Gonçalves, and Leonardo Rocha. 2024. Estratégias de Undersampling para Redução de Viés em Classificação de Texto Baseada em Transformers. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 144–152. <https://doi.org/10.5753/webmedia.2024.241229>
- [7] Tiago Gomes, André Regino, Rodrigo Caus, Victor Sotelo, and Julio Reis. 2025. Multi-Agent LLM Approach for Moderating E-Commerce Customer Service Responses. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 349–357. <https://doi.org/10.5753/webmedia.2025.16162>
- [8] Google. 2025. *Year in Search 2025: Brazil*. Google Trends. <https://trends.withgoogle.com/pt-br/year-in-search/2025/br/> Google Year in Search 2025 for Brazil; accessed: 2026-05-29.
- [9] Sayash Kapoor, Benedikt Stroebel, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI Agents That Matter. arXiv:2407.01502 [cs.LG] <https://arxiv.org/abs/2407.01502>
- [10] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI '15*). Association for Computing Machinery, New York, NY, USA, 3819–3828. <https://doi.org/10.1145/2702123.2702520>
- [11] LangGraph. 2025. LangGraph. <https://github.com/langchain-ai/langgraph>
- [12] Anthony Lee. 2026. *How AI Platforms Search: Fan-Out Query Behavior Across Intent Types, Verticals, and Platforms*. <https://doi.org/10.2139/ssrn.6652059>
- [13] Aldo Lipani, Gizem Gezici, Yücel Saygin, and Emine Yilmaz. 2021. Evaluation Metrics for Measuring Bias in Search Engine Results. *Information Retrieval Journal* 24, 2 (2021), 85–113. <https://doi.org/10.1007/s10791-020-09386-w>
- [14] Gabriel Prenassi, Rodrigo Souza, Henrique Sekido, Guilherme Fonseca, Marcelo Manzato, and Leonardo Rocha. 2025. Calibração de Sistemas de Recomendação com LLMs: Otimização de Prompts para Balancear Precisão, Diversidade e Justiça. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 103–111. <https://doi.org/10.5753/webmedia.2025.15613>
- [15] Ránqia Research. 2026. *Quem Influencia o ChatGPT no Brasil?* GEO Research Paper Vol. 01, Ed. 001. Ránqia. <https://ranqia.ai>
- [16] Jonathan J Ross, Ekaterina Khramtsova, Anton van der Vegt, Bevan Koopman, and Guido Zuccon. 2025. RARR Unraveled: Component-Level Insights into Hallucination Detection and Mitigation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) (*SIGIR '25*). Association for Computing Machinery, New York, NY, USA, 3286–3295. <https://doi.org/10.1145/3726302.3730337>
- [17] Stuart Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice-Hall.
- [18] Vagner Santana, Reinaldo Ferraz, and Jeniffer Vieira de Deus. 2026. On the Limits for Web Accessibility Evaluation Using AI Agents. In *Proceedings of the 23rd International Web for All Conference (W4A '26)*. Association for Computing Machinery, New York, NY, USA, 187–197. <https://doi.org/10.1145/3800424.3800445>
- [19] Vagner Figueredo De Santana. 2024. Challenges and Opportunities for Responsible Prompting. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 592, 4 pages. <https://doi.org/10.1145/3613905.3636268>
- [20] Vagner Figueredo de Santana, Sara E Berger, Heloisa Candello, Tiago Machado, Cassia Sampaio Sanctos, Tianyu Su, and Lemara Williams. 2025. Responsible Prompting Recommendation: Fostering Responsible AI Practices in Prompting-Time. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 836, 30 pages. <https://doi.org/10.1145/3706598.3713365>
- [21] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. 2026. AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion* 126 (Feb. 2026), 103599. <https://doi.org/10.1016/j.infus.2025.103599>
- [22] Sander Schulhoff. 2026. *Learn Prompting: Your Guide to Communicating with AI*. Learn Prompting. <https://learnprompting.org/> Open-source educational platform on prompt engineering and generative AI.
- [23] SearXNG Team. 2026. *SearXNG: A Privacy-Respecting Internet Metasearch Engine*. <https://docs.searxng.org/> Free and open-source metasearch engine; accessed 2026-06-03.
- [24] Cole Stryker. [n. d.]. Types of AI agents. <https://www.ibm.com/think/topics/ai-agent-types>
- [25] Masaki Suzuki and Yusuke Yamamoto. 2021. Characterizing the influence of confirmation bias on web search behavior. *Frontiers in psychology* 12 (2021), 771948.
- [26] Fons Wijnhoven and Jeanna Van Haren. 2021. Search engine gender bias. *Frontiers in big Data* 4 (2021), 622106.
- [27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2210.03629>

A GENERATIVE AI USAGE STATEMENT

No Generative AI (GenAI) was used in the preparation of this manuscript. GenAI was used though in a post-writing step to perform an initial ‘review’ and recommend a bullet list of improvements before the submission. Prompt:

Create an academic paper review for the attached PDF file. Consider the criteria in the following CFP: [URL]. Present your feedback in a bullet list, reduced format.

B ETHICAL CONSIDERATIONS

Considering the applicable Brazilian norms and guidelines (*Resolução CNS 466/2012*, *Norma Operacional 001/2013*, *Resolução CNS 510/2016*, and *Resolução CNS 674/2022*), this study operates within the domain of research using publicly available, which exempts this study from Institutional Review Board (IRB) review.