

Supplementary material : Keyword search is all you need: Achieving RAG-Level Performance without vector databases using agentic tool use

Anonymous submission

Abstract

While Retrieval-Augmented Generation (RAG) has proven effective for generating accurate, context-aware responses based on existing knowledge bases, it presents several challenges including retrieval quality dependencies and integration complexity. Recent advances in agentic-RAG and tool-augmented LLM architectures have introduced alternative approaches to information retrieval and processing. We question how much additional value vector databases and semantic search bring to RAG over simple, agentic keyword search in documents for question-answering. In this study, we conducted a systematic comparison between RAG-based systems and tool-augmented LLM agents, specifically evaluating their retrieval mechanisms and response quality when the agent only has access to keyword search tools. Our empirical analysis demonstrates that tool-based keyword search implementations within an agentic framework can attain over 90% of the performance metrics compared to traditional RAG systems without using a standing vector database. Furthermore, the tool-augmented approach offers distinct advantages in terms of system adaptability and lower maintenance overhead, particularly in scenarios requiring frequent updates to knowledge bases. We additionally explore preliminary results using Claude's Computer Use capability as an alternative implementation approach. This work contributes to the growing body of knowledge on optimal architectures for information retrieval and response generation in LLM-based systems.

Supplementary material

Agent terminal tool instructions

We use the standard zero shot agent template from langchain but provide a detailed description of the following tools it can use via the terminal.

```
1 terminal: Run shell commands on this
    Linux machine to search
    information in the "files/"
    folder. The commands to use are:
2
3
4 ```pdfmetadata.sh```
5 -----
6 YOU MUST first print details of pdf
    files in the files/ folder (ALWAYS
    start with this without any changes)
7 # sh pdfmetadata.sh
8
```

```
9 This gives you file level metadata that
    is useful to narrow down the search.
    Then use rga or pdfgrep. Action input
    must start with rga or pdfgrep and
    contain the full command.:
10
11 ```rga```
12 -----
13 A command line tool to search through
    files via keyword searches and regex
    patterns. All files relevant to this
    task are in the files/ folder.
14
15 - To find a search term in specific file
    (use regex pattern)
16 rga 'searchterm\w*' ./files/filename.pdf
17
18 - To search with multiple keyword
    matches across multiple files:
19 rga 'keyword1|keyword2|keyword3' ./files
    /
20
21 - Use -i for case insensitive search.
22
23
24 ```pdfgrep```
25 -----
26 Another commandline tool specifically
    for search with PDFs. Useful for
    special cases when:
27
28 - search in a folder with pdfs, across a
    specific page range for one or more
    keywords (-i is case insensitive, -n
    includes page numbers in output, -r
    is recursive search in files folder,
    and -P is perl compatible regex).
    With pdfgrep you must include '(' ')'
    brackets for the pattern
29 pdfgrep -inrP --page-range 1-4 '(
    keyword1|keyword2)' ./files/
30
31 - Search all .pdf files whose names
    begin with foo recursively in the
    current directory:
32 pdfgrep -r --include "foo*.pdf" pattern
33
34 Other tips:
```

```

35 *****
36 - If a complex query fails, try a series
    of simpler queries instead.
37 - ALWAYS try to return larger context
    with -C 5 with both rga nad pdfgrep
    to get 2 or more lines around the
    returned keyword match.
38 - remember that you MUST do 'sh
    pdfmetadata.sh' first to understand
    what files you are dealing with and
    then continue your search.
39 - your action should always just be "
    terminal" and action input is the
    full command you want to run in the
    terminal
40 *****

```

```

29     END {
30         print file, title, author, subject
            , keywords, creator, producer,
            creationdate, moddate, tagged,
            pages, encrypted, pagesize,
            filesize, optimized, pdfversion
31     }'
32 done

```

PDF metadata extraction script

While 'rga' and 'pdgrep' are tools availalbe in the open source, we created a short script to extract metadata from PDFs inside a directory in a structured way; this shell script is shown below:

```

1     #!/bin/bash
2
3     # Directory containing PDF files
4     directory="./files/"
5
6     # Print headers
7     echo -e "File\tTitle\tAuthor\tSubject\t
            tKeywords\tCreator\tProducer\t
            tCreationDate\ttModDate\ttTagged\ttPages
            \tEncrypted\ttPageSize\ttFileSize\t
            tOptimized\ttPDFVersion"
8
9     # Find PDF files and extract metadata,
    ignoring errors
10    find "$directory" -name "*.pdf" | while
        read -r file; do
11        # Extract metadata using pdftinfo,
            ignoring errors
12        pdftinfo "$file" 2>/dev/null | awk -v
            file="$file" '
13        BEGIN { FS = ": "; OFS = "\t"; title
            =""; author=""; subject="";
            keywords=""; creator=""; producer
            =""; creationdate=""; moddate="";
            tagged=""; pages=""; encrypted
            =""; pagesize=""; filesize="";
            optimized=""; pdfversion="" }
14        /Title:/ { title=$2 }
15        /Author:/ { author=$2 }
16        /Subject:/ { subject=$2 }
17        /Keywords:/ { keywords=$2 }
18        /Creator:/ { creator=$2 }
19        /Producer:/ { producer=$2 }
20        /CreationDate:/ { creationdate=$2 }
21        /ModDate:/ { moddate=$2 }
22        /Tagged:/ { tagged=$2 }
23        /Pages:/ { pages=$2 }
24        /Encrypted:/ { encrypted=$2 }
25        /Page size:/ { pagesize=$2 }
26        /File size:/ { filesize=$2 }
27        /Optimized:/ { optimized=$2 }
28        /PDF version:/ { pdfversion=$2 }

```

Example detailed agent run

The table below shows a detailed run of how our agent plans and executes commands to reach the final answer. For easier readability, commands run in the terminal corresponding to the original user question are shown in blue. The agent first uses the 'pdfmetadata' shell script to discover details about files existing in a local folder. Then, it uses 'rga' to look for content related to Hyperledgerfabric. Then, it dives deeper into specific pages using the 'pdfgrep' command to finally get the context required for the final answer.

Step	Agent observation
Question	What are the three main components in Hyperledger Fabric and what role does each component play in the system?
Action input	<code>sh pdfmetadata.sh</code>
Observation	-e File Title Author Subject Keywords Creator Producer CreationDate ModDate Tagged Pages Encrypted PageSize FileSize Optimized PDFVersion ./files/Blockchain-Solana.pdf Markus Richter Microsoft® Word 2019 Microsoft® Word 2019 Tue Jul 12 00:40:20 2022 UTC Tue Jul 12 00:40:20 2022 UTC yes 27 no 595.3 x 841.9 pts (A4) 594798 bytes no 1.7
Thought	The pdfmetadata output does not seem to contain information directly relevant to the main components of Hyperledger Fabric. I should search through the PDF files using the rga or pdfgrep tools to look for content related to "Hyperledger Fabric components", "Fabric architecture", or similar phrases.
Action input	<code>rga -i 'hyperledger fabric components fabric architecture' ./files/</code>
Observation	./files/BlockchainSolana.pdf:Page 14: reflected in the increase in performance and strength on confidentiality. The Hyperledger Fabric architecture is shown in Fig. 9. ./files/BlockchainSolana.pdf:Page 15: Fig. 9. Hyperledger Fabric Architecture
Observation	The search results indicate that the PDF file "BlockchainSolana.pdf" contains information about the Hyperledger Fabric architecture and its main components. I should skim through the relevant pages to identify the three main components and their roles.
Action input	<code>pdfgrep -inrP -page-range 14-16 '(component architecture)' ./files/BlockchainSolana.pdf -C 5</code>

Observation

14:Fig. 8. EOS Architecture 14- There is also rotation mechanism in the selection of producers: every three seconds, the 21 producers are selected from all producers again. This means producers 14-without enough computing power will be sifted out. With no peer competition and 14-shorter consensus confirmation time, it is possible for EOS to improve its scalability – ... **Context truncated** ... 15:Fig. 9. Hyperledger Fabric Architecture 15: There are three main components in Hyperledger Fabric: Membership, Blockchain 15-and Chaincode. Membership part provides identification services. Blockchain part 15-provides consensus services. Chaincode part is a program that acts as smart contracts in this system. In enterprise scenarios, each node could access this system through the membership services. The network is permissioned because the participants are known to each other, ra- – 16- Another attractive attribute of Fabric is the high confidentiality. The lack of confidentiality can be problematic for many enterprise-use cases, because it is impossible 16-to maintain business relationships in a completely transparent network. Hyperledger 16:Fabric enables confidentiality through its channel architecture and private data feature. The system could set the availability of specific data by assigning authorized peers. The assignment of confidentiality is shown in Fig. 11.

Final Answer

The three main components of Hyperledger Fabric are:

1. Membership - Provides identification services and allows nodes to access the system through membership services.
2. Blockchain - Provides consensus services for the distributed ledger.
3. Chaincode - Refers to the programs or smart contracts that execute on the blockchain network.

The membership component handles identity management and authentication of nodes. The blockchain component maintains the distributed ledger and ensures consensus among nodes through its consensus protocol. The chaincode component allows developers to write and deploy smart contracts that encapsulate the business logic and are executed on the blockchain network.

Table 1: Example agent agent run showing use of linux command line tools to answer the original user question.

Computer Use Agent Interactions

Below are example responses from our agent to various questions about American Express's 2022 financial statements, including the context used to derive each answer. These examples demonstrate our Computer Use agent's capabilities, which employs command-line tools and direct file operations to extract and verify information from PDF documents without relying on vector databases.

Question	Which debt securities are registered to trade on a national securities exchange under American Express' name as of 2022?
Output	As of 2022, American Express has only its Common Shares (par value \$0.20 per share) registered under Section 12(b) of the Securities Exchange Act to trade on the New York Stock Exchange under the symbol 'AXP'. No debt securities are registered to trade on any national securities exchange.
Context	The company explicitly states under 'Securities registered pursuant to Section 12(b) of the Act' only its common shares. While American Express has significant debt securities outstanding, these are not exchange-listed. The company maintains debt ratings from major agencies (Moody's, S&P, and Fitch) for its unsecured debt programs. Their debt securities are primarily issued through private/institutional channels rather than listed exchanges.
Question	What are the geographies that American Express primarily operates in as of 2022?
Output	American Express operates globally with a strong presence in the United States and significant international operations. As of 2022, approximately 78% of their total revenues net of interest expense were generated from U.S. operations, while 22% came from international activities. Their largest international presence outside the United States includes the United Kingdom, Japan, the European Union, Australia, Canada, and Mexico.
Context	The company operates in approximately 103 countries and territories through network partnerships. They have a regional structure divided into: United States, EMEA (Europe, Middle East and Africa), APAC (Asia Pacific, Australia and New Zealand), and LACC (Latin America, Canada and the Caribbean). International operations are managed through their International Card Services (ICS) segment.

Table 2: Computer Use Agent Query Examples - Part 1

Question	Does AMEX have an improving operating margin profile as of 2022?
Output	Operating margin is not a particularly useful or relevant metric for analyzing American Express's performance. As a financial services company, AMEX's profitability is better measured through metrics like Return on Equity (ROE), net interest margin, and efficiency ratios that are more appropriate for financial institutions.
Context	The 10-K does not discuss or report operating margin as a key metric. AMEX focuses on financial services-specific metrics like ROE, which is defined as 'net income divided by average shareholders' equity'. The company emphasizes credit performance, loan loss reserves, and net interest income as key performance indicators.
Question	What drove gross margin change as of the FY2022 for American Express?
Output	Gross margin is not a relevant metric for American Express as a financial services company. Instead, key profitability metrics include net interest income, net interest yield, and net interest margin, which are more appropriate for analyzing financial institutions. For FY2022, American Express reported that their net interest income increased 28% year-over-year.
Context	No mentions of 'gross margin' were found in the 10-K, as this metric is more relevant for companies selling physical goods. Net interest income increased to \$9.895 billion in 2022, up 28% from 2021. Net interest yield is a key metric that measures profitability of the card member loan portfolio.
Question	How much has the effective tax rate of American Express changed between FY2021 and FY2022?
Output	American Express's effective tax rate decreased by 3 percentage points, from 24.6% in FY2021 to 21.6% in FY2022.
Context	The reduction in the effective tax rate was primarily due to discrete tax benefits in 2022. These benefits were related to the resolution of prior-year tax items. Tax rates in both years were influenced by pretax income levels and geographic business mix.

Table 3: Computer Use Agent Query Examples - Part 2

Question	What was the largest liability in American Express's Balance Sheet in 2022?
Output	The largest liability on American Express's Balance Sheet in 2022 was Customer deposits at 110,239 million (or 110.2 billion).
Context	Found in the Consolidated Balance Sheet on page 95 of the 10-K. Customer deposits significantly exceeded other major liability categories at 110,239 million, compared to: Accounts payable (12,133 million), Long-term debt (42,573 million), and Other liabilities (37,350 million).
Question	Was American Express able to retain card members during 2022?
Output	Yes, American Express was successful in retaining card members during 2022. The company explicitly stated that 'Card Member retention remained high' during 2022, which contributed to a 17% year-over-year increase in net card fees.
Context	Card member retention success was evidenced by record levels of new card acquisitions combined with high retention rates. Net card fees increased 17% year-over-year, demonstrating the effectiveness of their premium value propositions. The company's investments in card member value and benefits appeared to support strong retention.

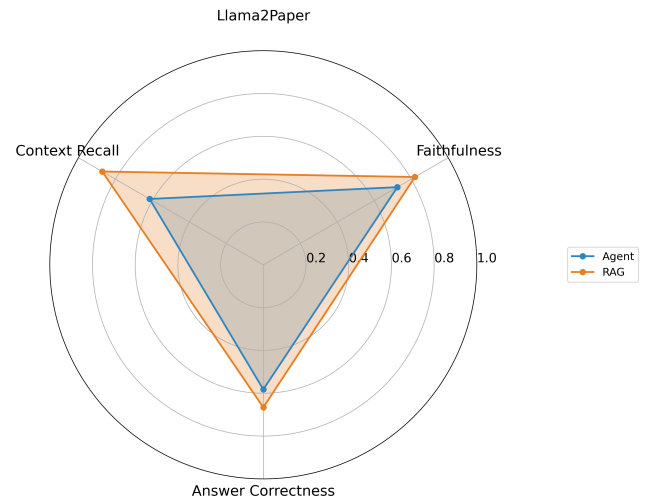


Figure 2: Coverage comparison of Tool-Augmented Agent vs RAG metrics for Llama2 paper dataset

Table 4: Computer Use Agent Query Examples - Part 3

Metrics Coverage Across All Datasets

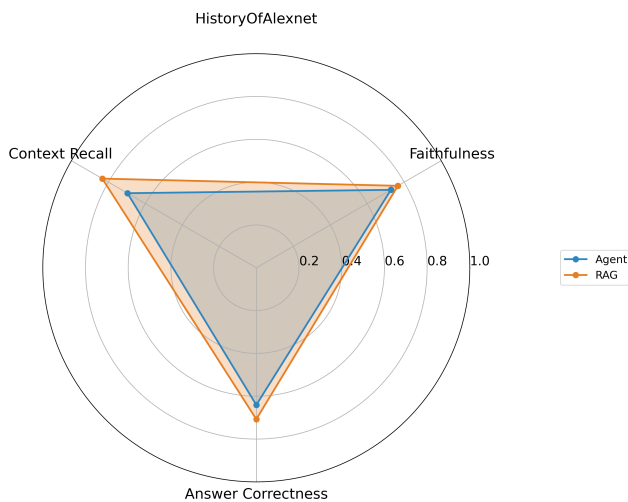


Figure 1: Coverage comparison of Tool-Augmented Agent vs RAG metrics for History of AlexNet dataset

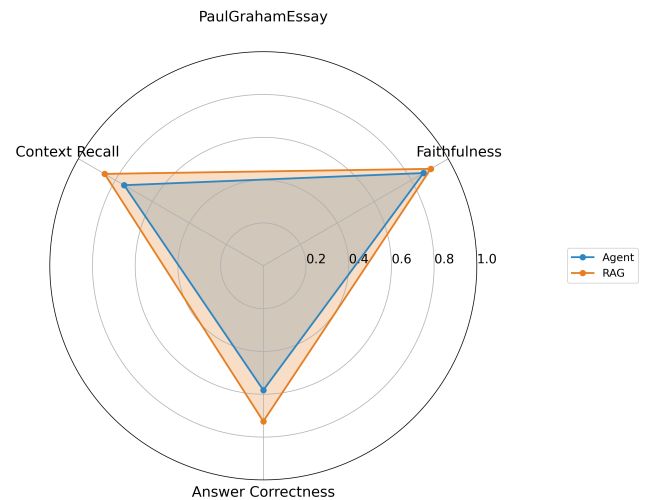


Figure 3: Coverage comparison of Tool-Augmented Agent vs RAG metrics for Paul Graham Essay dataset