



RETRIEVAL AUGMENTED GENERATION FOR OPERATIONS ORDER EVALUATION

Presenter: Wilmer Maldonado
Project Members: Wilmer Maldonado, Dr. Dan Cassenti, Brian Burbank

18 SEPTEMBER 2025

Distribution Statement:

Approved for Public Release

POC:

wilmer.maldonado.ctr@army.mil

AGENDA



| ORDER: | INFORMATION: |
|-----------|---|
| 01 | TOPIC: BACKGROUND AND PROBLEM |
| 02 | TOPIC: OBJECTIVES |
| 03 | TOPIC: DATASET |
| 04 | TOPIC: METHOD - RAG WITH OUTLINE SCHEMAS |
| 05 | TOPIC: LLM APP TOOLS |
| 06 | TOPIC: EVALUATION |
| 07 | TOPIC: LIMITATIONS |
| 08 | TOPIC: QUESTIONS |

COMMAND AND CONTROL (C2) AND MDMP



BACKGROUND

- Command and Control (C2) defines how DoD makes operational decisions.
- C2 requires commanders to review situational reports (SITREPs) and select Course of Actions (COAs).
- For higher-echelon (Battalion and above) the 7-step Military Decision Making Process (MDMP) is utilized to select COAs.
- Army planners employ the MDMP and produce documents that detail how to execute the approved COA.

Step 1: Receipt of Mission

Step 2: Mission Analysis

Step 3: COA Development

Step 4: COA Analysis

Step 5: COA Comparison

Step 6 COA Approval

Step 7: Orders Production, Dissemination, and Transition

BASIC OPORD IN “SMESC”/”SMEAC” FORMAT



MAIN OUTLINE

1. **Situation**: The General is coming, and I want a flag flying off a flagpole as soon as possible.
2. **Mission**: Company A will utilize the materials listed below to erect a flagpole in front of the Company headquarters tomorrow.
3. **Execution**:
 - a. **Commander’s Intent**. Honor the nation by having a flag flying from a well-secured pole before the General arrives in two days.
 - b. **Tasks**.
 - (1) First Platoon will dig a four-foot-deep hole using company tools centered and 20 feet in front of HQ.
 - (2) Second Platoon will mix concrete using a wheelbarrow, cement, sand, and gravel and pour it into the hole.
 - (3) Third Platoon will install the flagpole and rig it upright and plumb.
 - c. **Coordinating Instructions**. Platoon Leaders will cooperate to make sure the work is properly sequenced.
4. **Sustainment (Administration and Logistics)**:
 - a. **Administration**. Platoon Leaders will check with First Sergeant to make sure no work party troops are on light duty.
 - b. **Logistics**. All project materials and gloves are staged in the company storage building. Fill cooler and keep troops hydrated.
5. **Command and Signal**:
 - a. **Command**. Company Commander is out tomorrow. The Executive Officer is in charge.
 - b. **Control**. The XO will inspect pole emplacement while cement is still wet.
 - c. **Signal**. Use cell telephones as needed.

CHALLENGES IN GENERATING OPERATION ORDERS



HOW ARE OPORDS GENERATED?

- OPORDs are manually produced and evaluated by trained MDMP planners.
- Planners compile the OPORD from information extracted from various data sources using Microsoft Word.



CHALLENGES

- Writing an OPORD is a time-sensitive and precise process.
- Due to complexity and information-dense reports, mistakes and errors of omission can occur.
- Commanders sometimes must carry on with the missing information or request that the MDMP team revise.



HOW CAN THE PROCESS BE MADE EASIER?

- Augment the human team with artificial intelligence (AI) to help generate, understand, and evaluate OPORDs.

OBJECTIVES



KEY OBJECTIVES

- Reduce the cognitive load of commanders during OPORD generation and evaluation.
- Help Army OPORD Planners catch mistakes, such as errors of omission.
- Increase operational efficiency by reducing time-wasteful iterative revisions.
- Achieve decision-dominance with Human and AI working together.

PROPOSED SOLUTION

- An app-based suite of autonomous agents and over-the-shoulder Large Language Model (LLM)-based OPORD analysis tools.
 - Two OPORD Question-Answering Modules: RAG QA and Feature Extractor QA
 - Two OPORD Evaluation Focused Modules: OPORD Evaluator and OPORD Task Tracker

DATASET



SOURCE

- Exercise Support Application (ESA) can be accessed with CAC credentials.
 - Training exercise database
 - Includes an exercise that holds documents produced during MDMP, including an OPORD.



DESCRIPTION

- Adheres to established doctrine in producing an OPORD consistent with field standards.
- Also includes OPLAN and its Annexes
 - The Annexes under the OPLAN follow the OPORD's SMESC formatting
 - Each Annex provides detailed guidance for its respective area (ex. Annex G - Engineering.)





METHOD: AUGMENTING THE KNOWLEDGE BASE OF AN LLM WITH RAG

RETRIEVAL AUGMENTED GENERATION

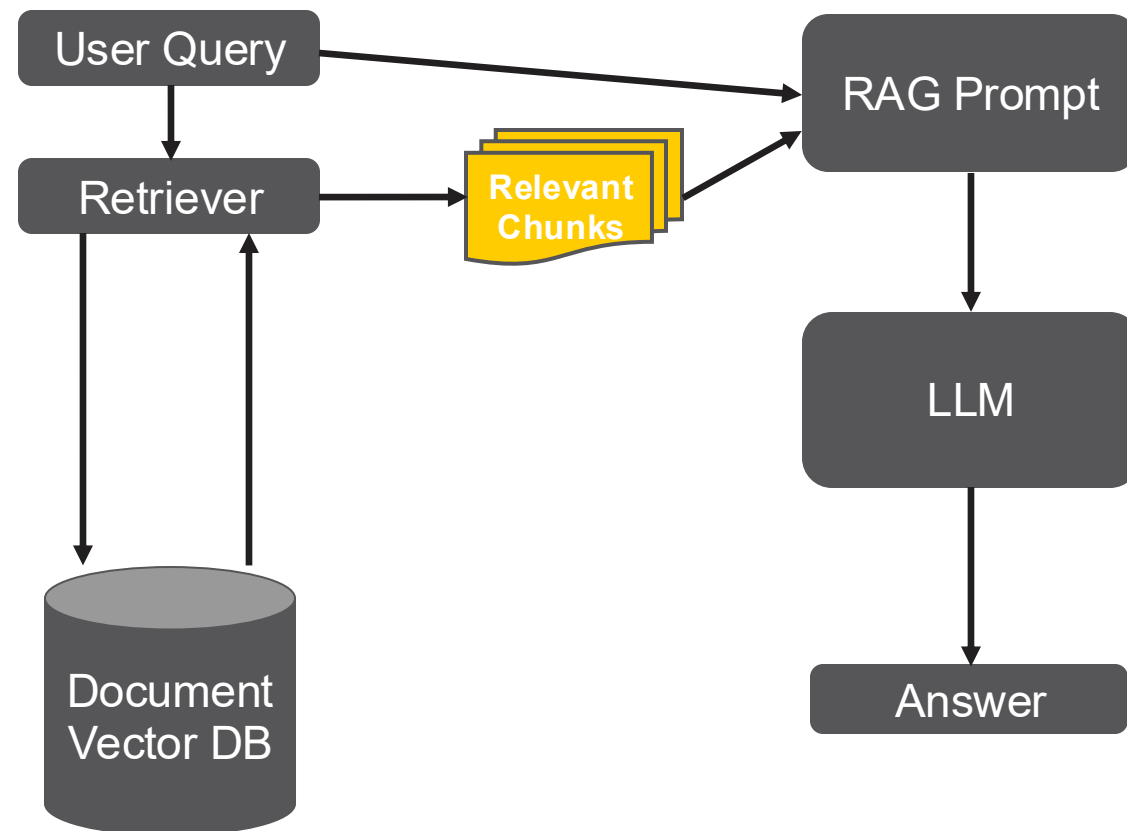
- Natural Language Processing (NLP) based method for sending relevant document sections or “chunks” to LLM for accurate question and answering.
- RAG Prompt is the final prompt sent to LLM – includes query, relevant chunks, and instruction.

RAG ADVANTAGES

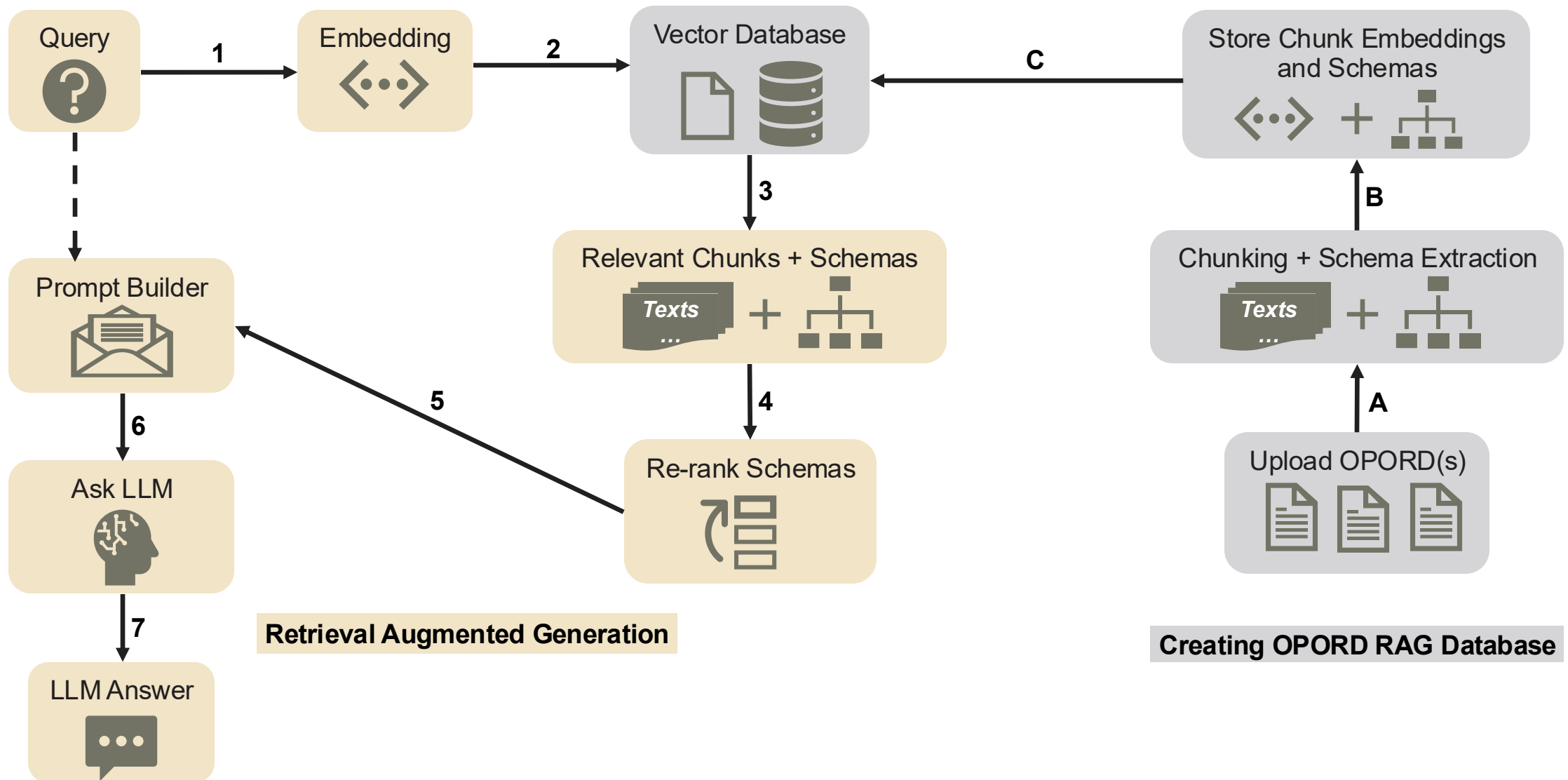
- No LLM model training required
- Low-maintenance when document database changes
- Updating backend to a newer LLM model or giving access to multiple is not an issue

HOW OUR APPROACH DIFFERS FROM TRADITIONAL RAG?

- Query-relevant document chunks are further extracted as schemas before retrieved and passed to LLM for answer generation.
- The schemas preserve the hierarchal structure present in the OPORD outline.



RAG WITH OUTLINE SCHEMAS

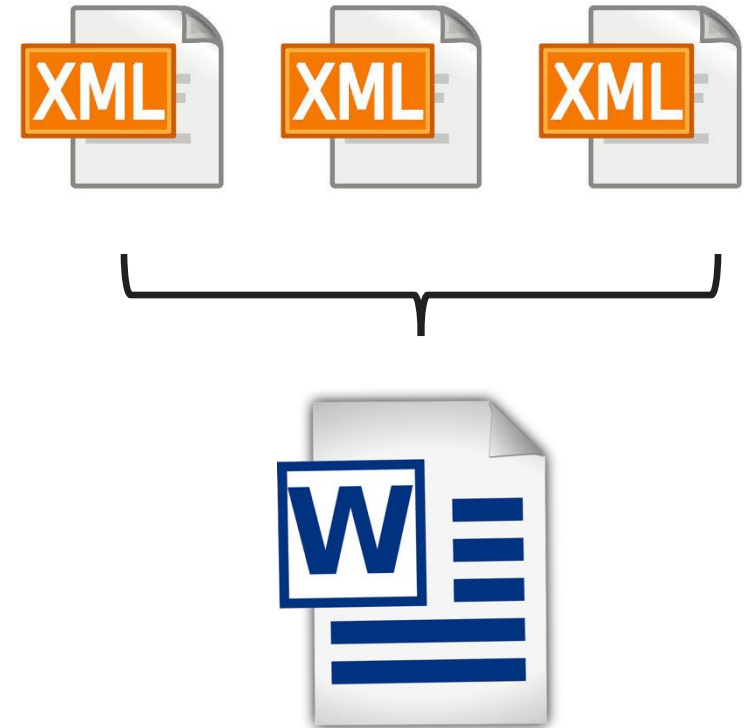


SCHEMA EXTRACTION



XML AND MICROSOFT WORD

- Army planners use Microsoft Word to create OPORDs.
- Microsoft Word uses XML (Extensible Markup Language) as the foundation of its modern file format since 2007 version.
- Each OPORD document chunk schema contains the following XML extracted elements:
 - Outline Indentation levels
 - Outline Notation/Markings Style
 - Tables
 - Figures
 - Document text and formatting
- Leveraged lxml, a xml parser library, to create the schemas
- lxml is python binded to C code libraries that are highly optimized and stable.
 - Processing time for ten OPORDs at four pages each is ~ 0.7s total.



LLM APPS: QUESTION AND ANSWERING



OPORD RAG QA

KEY FEATURES

- Chat-based RAG Question and Answering
- View a query's top ranked schemas/context
- OPOD filename filtering
- LLM Answer with reasoning explanation

PURPOSE

- Reduce cognitive load
- Quickly locate information
- Build human trust with answer transparency

OPORD Feature Extractor QA

KEY FEATURES

- Select Specific OPOD features to query
- Faster response times
- Low token usage / computational cost
- LLM Answer with reasoning explanation

PURPOSE

- Optimize query precision
- Planners can narrow their focus to key OPOD sections.

RETRIEVAL EXAMPLE

BASELINE RAG

Q: “Can you name all the general tasks under Execution Section?”

Chunk – Rank 1

“Purpose. The Force will execute the training schedule at FPVA to ...

Key Tasks.

....

Tasks.

General Tasks.

Deploy organic equipment, prime movers and MCPs for MUTA. This is an all-hands training evolution. ALL enlisted personnel and Officers will travel”

Chunk – Rank 2

“Oversea Force support, including hydration, messing, training material support, and medical support IAW Annex F. G6.

....

IAW Annex C, plan and execute support for and evaluation , if directed by

....

Packs (TACPAK (or PITTPAK) tasks on a go/no go basis.

...

SSB Data. ...



OUTLINE SCHEMA RAG

Q: “Can you name all the general tasks under Execution Section?”

Top-Ranked Schema

b. (U) Tasks.

(1) (U) General Tasks.

(a) (U) Deploy organic equipment, prime movers and MCPs for MUTA. This is an all-hands training evolution. ALL enlisted personnel and Officers will travel to FPVA to participate in training and evaluations, unless specifically excused by their senior FORHQ or MSC leader.

(b) (U) All echelon commanders (Cdrs) will ensure their commands provide personnel support to MUTA training. Personnel accountability is key.

(c) (U) Ensure members meet appropriate seasonal uniform requirements. MSC junior leaders should inspect individual soldier equipment for serviceability and completeness (load bearing equipment (LBE), safety equipment, hydration, sunscreen, etc.)/ready for the mission.

(d) (U) All leaders will carefully read this OPORD and communicate it to the lowest level within their respective units as a basic leadership measure.

(e) (U) Read Coordinating Instructions. Per Annex C, make movement plans and maintain positive control of personnel during movement to FPVA AO and re-deployment to home station (HS).

(f) (U) Conduct accountable equipment inventory upon end of mission and report mission complete as further described below.

ANSWER EXAMPLE

BASELINE RAG

Q: "Can you name all the general tasks under Execution Section?"

ANSWER

LLM Answer: The general tasks under execution section include the following:

1. Deploy organic equipment, prime ... senior FORHQ or MSC leader.
2. Provide Overseas force support, including hydration, messing, training material support, and medical support in accordance with Annex F.
3. Plan and execute support ... on a go/no go basis.
4. Ensure Statewide Agencies Radio System (STARS) ... MUTA Support.
5. Provide interest access information for FPVA.
6. Conduct monitoring for risks and problems at MUTA training sites.
7. Execute medical support as outlined in Annex F.
8. Conduct OCS (Officer Candidate School) training per Annex A.

Sources: (1)

Outside Context Note: No info outside provided text was used.

Confidence: 1.0



OUTLINE SCHEMA RAG

Q: "Can you name all the general tasks under Execution Section?"

ANSWER

LLM Answer: The general tasks under execution section include the following:

1. Deploy organic equipment, prime ... senior FORHQ or MSC leader.
2. All echelon commanders (Cdrs) will ensure ... is key.
3. Ensure members meet appropriate seasonal uniform requirements. MSC junior leaders should ... ready for the mission.
4. All leaders will carefully read this OPORD and ... as a basic leadership measure.
5. Read Coordinating Instructions. Per Annex C, make ... movement to FPVA AO and re-deployment to home station (HS).
6. Conduct accountable equipment inventory upon ... as further described below.

Sources: "Execution", "General Tasks"

Outside Context Note: No info outside provided text was used.

Confidence: 1.0

LLM APPS: OPORD TASK EVALUATION



OPORD Evaluator

KEY FEATURES

- Grades OPORD sections based on user-defined task criteria
- Detailed LLM evaluation explanation
- Utilizes METL Grading Scale:
 - **Trained (T)**: Highest grade / task complete
 - **Needs Practice (P)**: Passing grade / task satisfactory
 - **Untrained (U)**: Unsatisfactory grade / task incomplete

PURPOSE

- Reduce number of OPORD iterative revisions
- Simplify the OPORD Evaluation Process
- Reduce memory load on Army Planners

OPORD Task Tracker

KEY FEATURES

- Human-in-the-Loop Evaluation
- Task tracking and monitoring
- Custom task creation
- Detailed LLM evaluation explanation

PURPOSE

- Help Army OPORD Planners stay organized and track tasks
- Create OPORD evaluation dataset with logged human feedback

OPORD EVAL RESPONSE EXAMPLE



OPORD w/ all Task sections present.

Task: OPORD contains a complete "Tasks" section where General Tasks are listed. They should be separate from "Key Tasks", both should be under Execution Section.

LLM EVALUATION

LLM Eval Reasoning: The OPORD Execution Section successfully delineates both "Key Tasks" and "General Tasks," fulfilling the requirement for a complete "Tasks" section. The Commander's Intent clearly outlines the purpose and end state, while the General Tasks provide specific actions for personnel and equipment management. Each task is actionable and relevant to the overall mission, ensuring clarity and accountability. Overall, the section meets the evaluation criteria with no significant omissions.

LLM Eval Grade: T

Outside Context Note: No outside context was used in the evaluation.

Confidence: 0.95

OPORD w/ missing "Key Tasks" section

Task: OPORD contains a complete "Tasks" section where General Tasks are listed. They should be separate from "Key Tasks", both should be under Execution Section.

LLM Evaluation

LLM Eval Reasoning: The Execution Section of the OPORD provides a comprehensive list of General Tasks that are necessary for the training evolution at FPVA. However, it does not explicitly separate the General Tasks from any Key Tasks, which is a requirement per the evaluation criteria. While the tasks listed are relevant and detailed, the lack of a distinct separation between General and Key Tasks indicates that the section does not fully meet the specified standards. Therefore, it requires improvement to align with the evaluation criteria.

LLM Eval Grade: U

Outside Context Note: No outside context was used; the evaluation was based solely on the provided OPORD section and evaluation criteria.

Confidence: 0.85

EVALUATION



QUESTION AND ANSWERING

- **Single Document Questions**
 - Fact-seeking questions found in a single OPORD or Annex.
- **Multi Document Questions**
 - Fact-seeking questions found in a two or more OPORDs or Annexes.
- **Structural Questions**
 - Accurate answers depend on understanding the OPORD's formatting, such as section hierarchies/indentations and outline notation.
- **Table Queries**
 - Require looking up specific information from tables within the document(s).
- **Reasoning/Inference Queries**
 - Require deducing information that is implied but not explicitly stated.
- **10 Negative Questions**
 - LLM acknowledges when the document does not contain the answer.

| Question Category | Total # of Questions | Correct (%) |
|---------------------|----------------------|-------------|
| Single Document | 100 | 100 |
| Multi-Document | 40 | 95 |
| Structural/Outline | 40 | 100 |
| Table | 20 | 100 |
| Reasoning/Inference | 20 | 90 |
| Negative | 20 | 90 |
| Overall | 240 | 97.5 |

LIMITATIONS AND FUTURE WORK



LIMITATIONS

- Dataset is limited to one main OPORD exercise
- Not tested with multiple army planners for real training scenarios
- Question and answer (QA) evaluation dataset limited to 5 annexes
- Evaluation LLM system instruction has no real grading examples

FUTURE WORK

- Benchmark against baseline RAG and measure other advanced RAG evaluation metrics.
- Gather app tools feedback from trainee and former experienced OPORD Planners.
- Test a completely air gapped version of the app with a small local LLM and measure performance differences.

REFERENCES



- [1] J. J. Marr, *The military decision making process: Making better decisions versus making decisions better*, School of Advanced Military Studies, U.S. Army Command and General Staff College, 2001.
- [2] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang, "From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr. 2020, pp. 1–6.
- [3] Army Training and Doctrine Command, "Exercise Support Application". [Online]. Available: <https://be.tradoc.army.mil/esa-exercise-support-application>. [Accessed: 18-June-2025].
- [4] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, et al., "Active retrieval augmented generation," in *Proc. 2023 Conf. Empirical Methods Natural Language Processing*, 2023, pp. 7969-7992.
- [5] Y. Hu, Z. Lei, Z. Zhang, B. Pan, C. Ling, and L. Zhao, "GRAG: Graph Retrieval-Augmented Generation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, Apr. 2025, pp. 4145–4157. doi: 10.18653/v1/2025.findings-naacl.232.
- [6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. 36th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, 2022, pp. 1800–1813.
- [7] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 159:1–159:25.
- [8] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-bench and Chatbot Arena," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, 2023, Art. no. 2020, pp. 1–29.
- [9] S. Farquhar, J. Kossen, L. Kuhn, et al., "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, pp. 625–630, 2024. doi: 10.1038/s41586-024-07421-0.
- [10] P. Manakul, A. Liusie, and M. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023, pp. 9004–9017, doi: 10.18653/v1/2023.emnlp-main.557. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.557>. [Accessed: 18-June-2025].

THANK YOU.



U.S. ARMY