LLM Augmentations to support Analytical Reasoning over Multiple Documents

Raquib Bin Yousuf, Nicholas Defelice, Mandar Sharma, Shengzhe Xu, Naren Ramakrishnan
Department of Computer Science, Virginia Tech, Arlington, VA
Email: raquib@vt.edu, naren@cs.vt.edu

***Abstract*—Building on their demonstrated ability to perform a variety of tasks, we investigate the application of large language models (LLMs) to enhance in-depth analytical reasoning within the context of intelligence analysis. Intelligence analysts typically work with massive dossiers to draw connections between seemingly unrelated entities, and uncover adversaries' plans and motives. We explore if and how LLMs can be helpful to analysts for this task and develop an architecture to augment the capabilities of an LLM with a memory module called dynamic evidence trees (DETs) to develop and track multiple investigation threads. Through extensive experiments on multiple datasets, we highlight how LLMs, as-is, are still inadequate to support intelligence analysts and offer recommendations to improve LLMs for such intricate reasoning applications.**

***Index Terms*—augmented LLM, intelligence analysis, analytical reasoning, retrieval and search, large language model.**

I. Introduction

The impressive generation capabilities of LLMs [1]–[3], along with their successful application in diverse areas [4]–[7] led us to explore their utility for intelligence analysis (IA). Key IA tasks involve uncovering plots from textual reports such that interventions can be made to prevent unfortunate events. This entails making connections between seemingly unrelated entities and events. This undertaking traditionally requires significant investments of time and effort from human analysts [8], [9]. Essentially, they are carrying out a "connecting the dots" task, where dots are the information bits involving different entities/events in their reports.

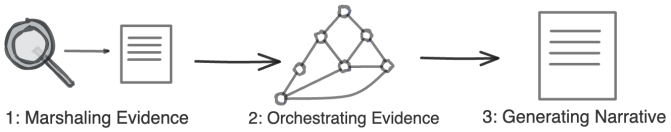


Fig. 1: Three steps to intelligence analysis (IA).

IA can be viewed as comprising the following subtasks: i) Marshaling evidence, ii) Orchestration of the gathered evidence: sensemaking and the construction of defensible and persuasive arguments from evidence, and iii) narrative generation. To be successful in IA, analysts often require an overarching capability of creative, speculative, and imaginative reasoning which helps build hypotheses. Even after making the connection between relevant information 'dots', an analyst must decide what this synthesized information portrays for the context at hand. New evidence can support or topple the hypothesis under consideration. After gathering all required evidence, analysts need to produce convincing arguments and reasoning to generate an appropriate alert to their superiors/authorities.

While numerous representations have been proposed for reasoning with LLMs [10]–[12] they are typically focused on individual questions or tasks [2], [13], [14]. Our focus here is on reasoning emergent from assimilating hundreds of documents, sometimes beyond context window limits for LLMs. We design a set of experiments to study whether LLMs can solve IA problems and if not, how can they be augmented to support such analysis. Our codebase is publicly available at https://github.com/DiscoveryAnalyticsCenter/speculatores. Our key contributions are:

1) We conduct the first investigation of the feasibility of using LLMs in intelligence analysis where both evidence-based reasoning and analytical creativity is of utmost importance.
2) We develop a three-step augmentation to support the use of LLMs for IA: i) Dynamic Evidence Trees (DETs), a memory module to help organize evidences, ii) Data condensation via LLMs, and iii) an LLM-driven search and retrieval process.
3) Applying our framework on multiple IA datasets, we show that while our augmentations help orchestrate and improves narratives on large datasets, LLMs still lack the analytical creativity to craft convincing arguments.
4) We outline detailed recommendations for applying LLMs to IA tasks and specifically how they can serve as modules in specific subtasks such as evidence marshalling and narrative generation.

II. Methodology

We design our experiments with three intelligence analysis datasets. These datasets contain field reports of various events conducted by adversaries. Among these, some may be relevant and others are irrelevant. Analysts need to find pertinent information by isolating relevant reports and piecing the information together to speculate the main plot. We pose four research questions (RQ) to understand whether LLMs can be useful for the IA process:

- **RQ1:** *Can LLMs solve IA problems on their own?*
- **RQ2:** *Does augmentation help? If so what kinds of augmentation support the IA task?*

Accepted in IEEE Big Data 2024.

- **RQ3:** *What is the effect of temperature and allowable context size on the analytical skills of LLMs?*
- **RQ4:** *Where can LLMs contribute in the 3-step process outlined in section §1?*

We start with an overview of how the experiments are designed to capture the efficacy of LLMs in IA. The goal of intelligence analysis is not just to summarize the reports, but rather to find out the connections, along with the respective backstories about the adversaries. A major limitation for LLMs to achieve this is the context length and loss of attention with the increasing input size [15].

A. Preliminaries

Each IA dataset is a set of chronological reports $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ such that report r_{i-1} originates or was assigned before r_i. When there are multiple reports available on a single day, we can order them arbitrarily.

We also have a set of evidential information nuggets called 'dots' $\mathcal{E}_\mathcal{D} = \{e_{d1}, e_{d2}, \ldots, e_{dm}\}$. We can model each report r_i as one evidential dot e_{dmi} or an amalgamation of multiple evidential dots, i.e., $r_i = \{e_{d1}, e_{d2}, \ldots, e_{dm}\}$. Thus, we can represent each dataset as a tuple of $(\mathcal{R}, \mathcal{E}_\mathcal{D})$.

Conceptually there can be two types of information dots in IA: i) evidential dots e_{di}, which come directly from the intelligence reports; ii) hypothesis dots h_{di}, which are produced by synthesizing multiple evidential dots (and/or other hypothesis dots). Thus each hypothesis dot is modeled as $h_{di} = \{h_{d1}, h_{d2}, \ldots, h_{dk}\}, \ldots \{e_{d1}, e_{d2}, \ldots, e_{dl}\}$.

The goal is to use the tuple $(\mathcal{R}, \mathcal{E}_\mathcal{D})$ to create a set of hypothesis dots, i.e. $\mathcal{H}_\mathcal{D} = \{h_{d1}, h_{d2}, \ldots, h_{dn}\}$. In the most basic format, a hypothesis dot h_{di} can thought to be built with two evidential dots; $h_{di} = \{e_{dq}, e_{dp}\}$. Thus, the LLM autoregressively generates each token $h_{di,j}$ of a hypothesis dot h_{di} as:

$$\mathbb{P}(\mathbf{h}_{di}) = \mathbb{P}(h_{di,1}, h_{di,2}, ..., h_{di,n}) \quad (1)$$

$$\simeq \prod_{j=1}^{n} \mathbb{P}(h_{di,j} | h_{di,1}, h_{di,2}, ..., h_{di,j-1}) \quad (2)$$

We postulate that the autoregressive nature of LLMs will help model a report with evidential dot (or dots) as a whole, instead of modeling reports as an amalgamation of a set of entities. We note that each evidential dot e_{di} is comprised of a set of entities. The autoregressive nature of LLMs will keep the rich contexts of evidential dots intact and build up each hypothesis dot (h_{dj}). We are postulating that LLMs with their auto-regressive and generalization properties will be able to model dots as standalone artifacts and connect them when needed.

> Dr. Clark Adams, a Middle Eastern expert whose office is at the Pentagon, was last seen on 13 April, 2003 after he left his home at 1830hrs for the Home Depot store on Lee Highway in Merrifield, VA. Dr. Adams's family could give no explanation for his disappearance. The FBI is treating his disappearance as a possible case of abduction.

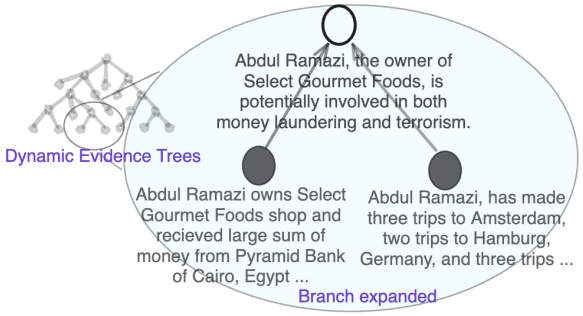
Fig. 2: Sample intelligence report

B. Initial experiments

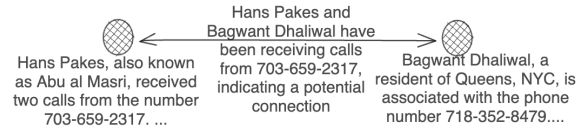
As the most basic step of the evaluation of LLMs for IA, we attempt to use each dataset in its entirety in the context windows, as the limit permits, to generate the narrative. We empirically test with varied system prompts. However, we find that the LLMs tend to only summarize the documents and present superficial information, despite numerous prompt variations to bring out implicit connections and the broader story. Moreover, the context length limits the number of documents that can be fit into the prompt. We randomly subsample the number of documents for datasets according to the context limit. We also quantitatively and qualitatively evaluate the responses as shown in the results section IV.

C. LLMs for Intelligence Analysis: Challenges

Through our initial experiments on LLMs without any augmentations, we identified some shortcomings for complex analytical reasoning tasks. A major limitation for LLMs is the context length and loss of attention with the increasing input size [15]. The two main challenges are i) lack of a proper memory module to keep track of all the evolving investigation threads, ii) limited context length that limits the number of reports that can be processed. To solve the first problem, we propose an augmented architecture with a much needed memory module to the pipeline, named "Dynamic Evidence Trees (DETs)". Memory modules are increasingly being used to augment LLMs in various applications [16]. As such, we devise a memory module in form of trees to help LLMs orchestrate evidence as the reasoning goes along. As an improvement on the second front, we also perform the tests with two different granularities of the reports. We propose condensing the reports into concise information chunks. In the later sections, we will formally define and present the inner workings of these two components.



(a) Regular DETs, dots: (black: document, white: hypothesis)



(b) Person-based DETs (shaded dots)

Fig. 3: Dynamic Evidence Trees

D. First augmentation: Dynamic Evidence Trees (DETs)

As an improvement to the basic LLM, we augment it with dynamic evidence trees (DETs). This is the main memory

component to save the evidence.

We define the main graph structure of DETs as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ denotes the vertices with $v \in \mathcal{V}$ such that each vertex v_i is a structure named DOT (i.e., information dot); and $\mathcal{E}$ denotes edges $e_{ij} \in \mathcal{E}$ which is the parent-children relationship between the DOT_i and DOT_j. Thus set of vertices $\mathcal{V}$ can be defined as $\{DOT_1, DOT_2, \ldots DOT_i\}$. Each DOT_i can be both evidential and hypothesis dots and we define DOT_i as a quadruple of $(information, DOT_{children}, DOT_{parents}, document)$, such that each DOT_i can have another set dots as children or parents. Evidential dots belong to one report of the dataset and we save that information in the node. DETs also holds a database object with LLM based embedding vectors for retrieval operation. Thus we denote DETs as tuple of $(\mathcal{DB}, \mathcal{G})$. During the creation of the DETs, we can think of each DOT as a subgraph with a tree like structure (with one node if it does not have any children or parents). LLM based operations will decide if multiple $\{DOT_1, DOT_2, \ldots DOT_i\}$ can be merged to create a new hypothesis DOT. This new DOT will have all the comprising dots as its children and the children dots will also save the new DOT as their parents. Thus, we are creating new connections among different disjoint subgraphs $\{DOT_1, DOT_2, \ldots DOT_i\}$ in the form of a new parent DOT.

We also experiment with a variant of DETs, where each dot is amalgamation of all information of a person in the dataset and the edges are defined as connections between two persons. Thus we can define DETs as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ denotes the vertices with $v \in \mathcal{V}$ such that each vertex v_i is information about a person; and $\mathcal{E}$ denotes edges $e_{ij} \in \mathcal{E}$ which is the connections between two persons, i.e., DOT_i and DOT_j.

With the help of DETs, LLMs try to build up hypotheses and investigation threads. Intelligence analysts go through a process of discovery and combining the dots to build up various hypotheses. Like human analysts might do, DETs helps to keep track of all the information dots and keeps connecting them with new relevant information in a tree like structure. Each hypothesis dot represents an investigation thread that may or may not get added to another related investigation thread. We report results for both regular and person-based DETs in section IV

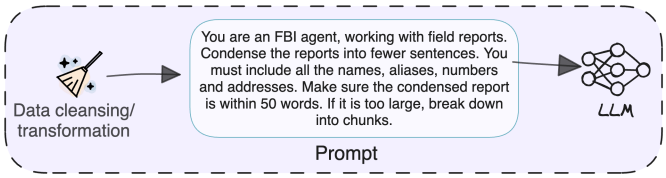


Fig. 4: Data condensation and dot extraction.

E. Second augmentation: Data Condensation

We utilize the language modeling capabilities of LLMs to digest the set of reports into usable information dots before generating reports. These condensed information dots can be merged with other information dots to create hypotheses. We empirically test various system prompts to break down the report in a zero-shot fashion such that a report r_i turns into an evidential dot e_{di}. We use dynamic pre-processing to break it down further, if deemed necessary by the LLM. All the reports in our experiments yielded a single dot each. This also indicates the concise reporting practice of the intelligence community.

F. Third augmentation: LLMs for Retrieval

To build up a DET, we require search and retrieval capabilities on the existing evidence list, given new evidence. We augment the LLMs with a two-step retrieval pipeline. The first step is powered by an LLM-based embedding vector database and similarity-based search. Following the recent use of LLMs-based embeddings for retrieval [17], [18], we adopt an instructor embedding model [19] for our database. The second step uses an LLM to filter the extracted set of DOTs.

G. The augmentations altogether

The augmented version of the experiment is designed to process intelligence report sequentially and keep track of the evidence in DETs. This sequential approach helps in three ways; first, to conform to the dynamic nature of how intelligence report are available and assigned over period of time; second, to help build up DETs to keep track of the emerging evidence; third, to get around the context limitation of the models. This also follows the natural tendencies of how analyst must make and keep the hypotheses running until enough evidence can be accrued and a reasoning can be established for each hypothesis. New reports and the resulting information can either approve or disprove any existing hypotheses. The DET-based augmented pipeline considers this aspect and inputs the reports iteratively to simulate the temporal continuum of intelligence analysis.

Each report goes through a pre-processing step based on the dataset provided and a set of information dots are extracted from the report. We initialize a database and keep updating it with new information dots during runtime. During the tree building operation, the system first attempts to search relevant DOTs from the DETs. The extracted $DOTs_{candidates}$ go through a parent level hypothesis DOTs identification and consolidation process. We take the lowest common parent hypothesis nodes and assign the new DOT to that node. This ensures that we are assigning the new DOT to the most relevant branch of the evidence tree. Afterwards, each evidential dots of the resulting new branch are extracted. LLMs will synthesize all the evidential dots to create a short narrative for this particular investigation thread. From the resulting DETs, we isolate the largest chain of events and use that as the main DET for the input. The full generated DETs along with its disjoint nodes also demonstrate how LLMs are being augmented to keep track different documents and how each of these documents are being utilized through the chain of events for building up to the final hypothesis. Fig. 5 shows the final steps of the proposed architecture and algorithm 1 and 2 shows the overall algorithm used for the build and merge operation.

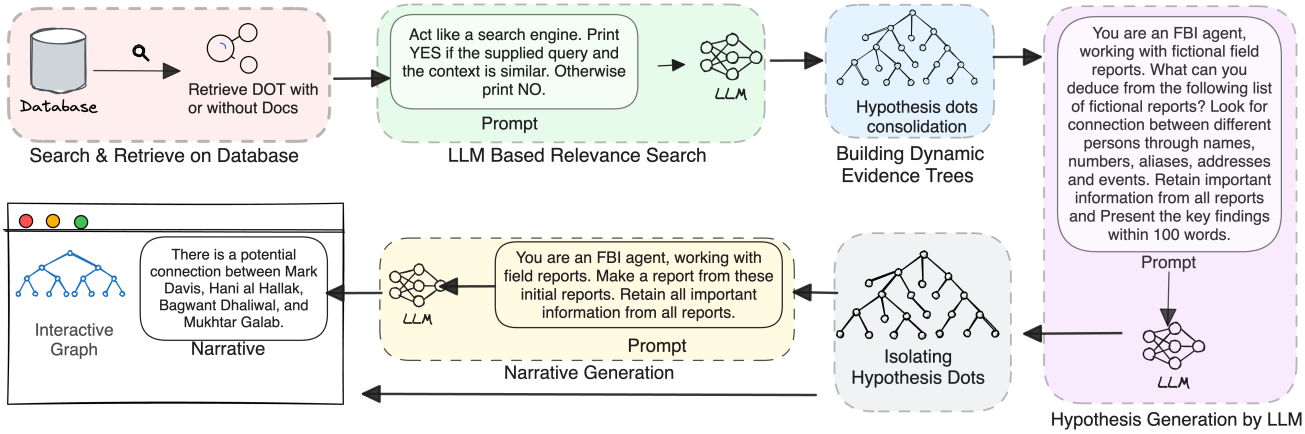


Fig. 5: All augmentations together with retrieve and merge

Algorithm 1 Build and Update DETs

Require: List of ordered reports, $\mathcal{R}$
Ensure: Dynamic Evidence Trees, DETs $(\mathcal{DB}, \mathcal{G})$
function CONDENSE AND EXTRACT DOTS(r_1)
  Invoke $\mathcal{LLM}(r_1)$
function SAVE INTO DATABASE(DOT)
  Invoke Embedding $\mathcal{LLM}$
  Save Embedding to DB
for $r \in \mathcal{R}$ **do**
  $DOTs$ = EXTRACT DOT(r)
  for $DOT \in DOTs$ **do**
    RETRIEVE AND MERGE2(DOT)
    SAVE INTO DATABASE(DOT)

Algorithm 2 Retrieve and Merge

Require: DOT
function RETRIEVE(DOT_{query})
  Vector Search in $DETs$
function LLM BASED FILTERING($DOT_{candidates}$, DOT_{query})
  Invoke $\mathcal{LLM}$
$DOT_{candidates} \leftarrow$ RETRIEVE(DOT)
$DOT_{filtered} \leftarrow$ LLM BASED FILTERING($DOT_{candidates}, DOT_{query}$)
if $|DOT_{filtered}| \neq \emptyset$ **then**
  $DOTs_{candidates} \leftarrow$ HYPO. NODE CONSOLIDATION($DOTs_{filtered}$)
  $DOTs_{candidates} \leftarrow$ EVID. NODE COLLECTION($DOTs_{candidates}$)
  $DOT_{new} \leftarrow$ Invoke $\mathcal{LLM}$
  RETRIEVE AND MERGE2(DOT_{new})
  SAVE INTO DATABASE(DOT_{new})

III. EVALUATION SETUP

In this section, we describe the layout of our evaluation procedure. Because we aim to test the efficacy of LLMs as an intelligence analyst, we evaluate the augmented architecture in an ablated fashion. We remove augmentations one by one and capture the improvement from the most basic bare-bone LLM. We also consider document-entity networks and clustering as a more traditional baseline sans any language modeling. For the basic form, we adopt two versions, one with a sub-sampled report set and another with the highest performing clusters on the report set coming from one of the traditional baselines. We use default parameters for the generation, with an enumerative testing with temperature and word limit to capture the randomness and creativity in reasoning. Our architecture is designed to use any type of LLM, both through API and local storage. It can fall back to a local LLM if it fails to get a results from the API calls. We experimented with five models from four different model families: i) GPT-3.5, and ii) GPT-4 from OpenAI, iii) Llama-2 [11] from Meta AI, iv) Mistral-7B [20] from MistralAI, and v) Gemma-2 [21] from Gemini platform, Google.

A. Datasets and ground truth

We utilize three datasets—Sign of the Crescent(Crescent), Atlantic Storm and Manpad—popular in training and analytics competitions [22]. The Crescent and Atlantic Storm datasets have their solutions divided into a few subplots and charts. Crescent has 3 subplots, each divided into 1-3 charts, with a total of 8 charts describing the information dots and the chain of reasoning. Similarly, the Atlantic Storm dataset has 6 subplots and 13 charts. All datasets use irrelevant reports as noise to make the problem more challenging. Statistics of the datasets and nodes created in DETs by the augmented architecture shown are shown in Table I. We differentiate surface level ground truth text with implicit ground truth. An AI model's true capabilities in case of solving IA process can only be evaluated by comparing the implicit ground truth. For Crescent and Atlantic Storm, we isolate the implicit information by manually going through the chart and removing the document level dots.

TABLE I: Dataset Statistics

Dataset	# Documents (Relevant/Irrelevant)	# Nodes Generated (Hypothesis/Evidential)
Crescent	41 (25/16)	45 (4/41)
Atlantic Storm	111 (65/46)	128 (17/111)
Manpad	50 (21/29)	50 (9/41)

B. Document-entity network and clustering

We consider a document-entity network as the basic baseline, as demonstrated in the entity graph based approaches [22]–[24]. These works typically consider the story as an established connection between a set of entities by graph properties. The task of sense-making, finding out the bigger story, and narrative generation are still up to the analysts.

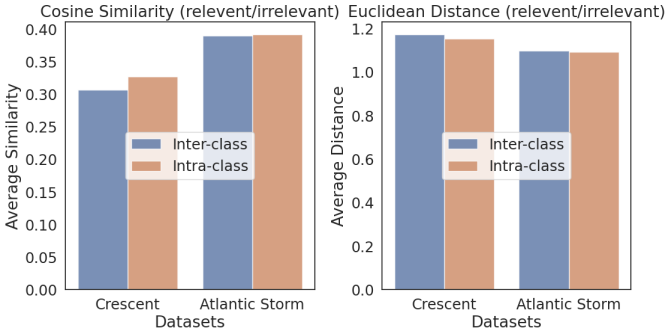


Fig. 6: Intra and inter class distances and similarity for relevant and irrelevant documents in datasets

Visual analytics tools were also developed to aid analysts with the sense-making task [25]–[27].

An objective of solving IA is to correctly isolate the irrelevant reports from the dataset. Traditional distance based methods, coupled with document embeddings, are often not enough to correctly classify the relevant reports. Fig. 6 shows how the inter and intra class distance are almost equal for relevant and irrelevant reports for Crescent and Atlantic Storm datasets. To classify the relevant and irrelevant datasets, we used multiple clustering algorithms [28], [29] on the embedding vectors and report the maximum score for each dataset in Table II.

C. Evaluation Metrics

We report the F1 score for the relevant/irrelevant document classification accuracy in all three datasets. We also compare the ground truth narrative with model narrative by adopting traditional ML metrics, i.e., average ROUGE 1/2/L(ROUGE-Lsum) [30], and METEOR [31]. To capture the quality of the narratives, we also employ GPT-4 [32] ratings. Recent works established LLMs as viable judge to evaluate the quality of open ended text generation [33]–[37] and to address the burdensome work of evaluation with human preferences [12], [38], [39]. We identified three important qualities for the model responses, i.e., relevance, coverage, and thoughtfulness, and test GPT-4 ratings in respect to these qualities.

TABLE II: Classification F-1 score on three datasets for LLM responses

Method	Crescent	Atlantic Storm	Manpad
Clustering	0.65	0.72	0.56
Basic Prompt	n/a	n/a	n/a
Basic Prompt (clustered)	0.65	0.72	0.56
Augmented w/o data condensation	0.80	**0.80**	0.64
Full augmentation (Regular DET)	**0.81**	0.78	**0.65**

IV. Results

Our experiments show that even with the augmented architecture, LLMs are unable to go beyond superficial reasoning. In this section, we expand on each RQ with the respective results.

A. RQ1 and RQ2: Can LLMs solve IA datasets on their own? Does augmentation help? If so what kind of augmentation helps?

To answer this set of RQs, we compare narratives from different experiments to the ground truth solutions. For each of Crescent and Atlantic Storm, we combined the ground truth of the charts into one report. For Manpad, we followed the rule-based grade to isolate relevant documents, with the entities extracted and desirable ones.

We first compare the classification accuracy for relevant documents. LLMs were only marginally better than clustering which suggests that the LLM-driven retrieval process is not upto mark, even though we had a two-step retrieval process.

Moreover, we also compare the narrative qualities against ground truth. The results for GPT-3.5 (Table IIIa) show that the basic prompted LLM has unsatisfactory metrics against the ground truth. Because our ground truth is the implicit text, rather than the superficial text, it shows that LLM was unable to pick up and guess implicit stories from the documents. We also noticed a trend of subpar scores in the Atlantic Storm dataset which indicates that LLMs struggle with larger datasets. To answer if augmentations help, we plot the normalized metrics for different methods with multiple temperatures (0.2-1.5) in Fig. 7. DET (regular) and DET (person-based) show improved performance across metrics and datasets. We also observe the largest improvement with the augmentation occurs for Atlantic Storm, which indicates that the augmented version helps in better structuring the large sets of reports. In other datasets however, it only slightly outperforms basic strategy. Moreover, the augmented version's performance degraded once the data condensation feature was turned off, which suggests that, to build a structured architecture, summarization and cleaning the data helps tremendously.

Traditional evaluation metrics often fail to correctly quantify the open-ended generation for the desirable characteristics [33], [40], so we also used GPT-4 to score the quality of the narratives. Across the qualities (i.e., relevance, coverage and thoughtfulness), we quantify the ratings in single Likert scale score (1-7) (Table IIIb). The trend confirms the narrative quality scores reported in Table IIIa. However, in the next section, we expand on a potential pitfall of traditional evaluation systems.

B. RQ3: What is the effect of temperature and allowable context size on the analytical skills of LLMs?

We also investigate if it is possible to increase the creativity/analytical skills by utilizing the randomness and creativity focused prompts for language models. For the basic setting, we tested two types of prompts with one containing system prompts about "being creative and imaginative" in reasoning. For the augmented architecture, the quality of the narrative is affected by both temperature and allowable context window limit for each hypothesis dot. Rather than limiting the generation by setting the token length, we asked the model to curb generation within a certain word limit by prompt. We carried out parameter sweep for this setup on Crescent dataset and

TABLE III: Evaluation of narrative generation.

(a) ROUGE and METEOR scores on three datasets for GPT-3.5 (temperature=0.7)

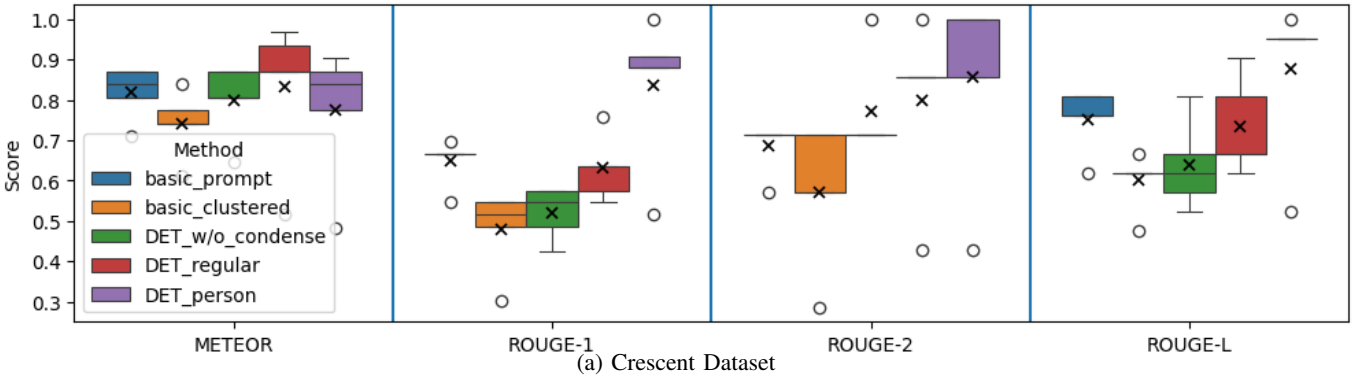
	Crescent				Atlantic Storm				Manpad			
Method	METEOR	R1	R2	RL	METEOR	R1	R2	RL	METEOR	R1	R2	RL
Basic Prompt	0.27	0.22	0.05	0.16	0.25	0.24	0.04	0.18	0.33	0.27	0.09	0.21
Basic Prompt (clustered)	0.26	0.18	0.05	0.14	0.21	0.21	0.03	0.15	**0.35**	0.35	**0.12**	0.26
DET (regular w/o data condensation)	0.25	0.19	0.05	0.13	0.20	0.17	0.03	0.13	0.20	0.20	0.03	0.14
DET (regular)	**0.29**	0.18	**0.07**	0.13	**0.34**	**0.29**	**0.10**	**0.21**	0.27	**0.36**	0.07	**0.28**
DET (person-based)	0.27	**0.29**	**0.07**	**0.20**	0.17	0.27	0.04	0.17	0.19	0.24	0.03	0.15

(b) GPT-4 Score on Likert chart (1-7) for difference quality of the narratives

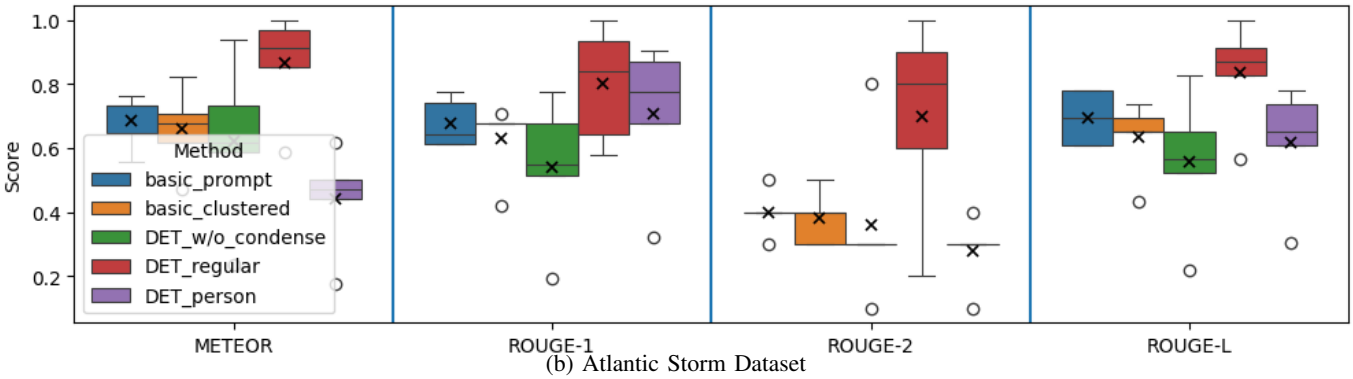
Dataset	Crescent			Atlantic Storm			Manpad		
Methods	**Relevance**	**Coverage**	**Thoughtfulness**	**Relevance**	**Coverage**	**Thoughtfulness**	**Relevance**	**Coverage**	**Thoughtfulness**
Fully augmented	5	3	5	2	2	5	1	1	4
Basic (sub-sampled)	3	2	3	1	1	3	3	2.5	5
Basic (clustered)	5	4	5	1	1	3	3	1	4

(c) ROUGE and METEOR scores on three datasets for different LLMs

	Crescent				**Atlantic Storm**				**Manpad**			
Model	METEOR	R1	R2	RL	METEOR	R1	R2	RL	METEOR	R1	R2	RL
GPT-4	0.25	0.23	0.07	0.15	0.23	0.23	0.04	0.16	0.28	0.31	0.07	0.19
Llama-2	0.20	0.19	0.03	0.11	0.21	0.21	0.03	0.15	0.28	0.36	0.08	0.26
Mistral-7B	0.26	0.23	0.07	0.16	0.25	0.20	0.04	0.15	0.34	0.37	0.11	0.26
Gemma-2	0.17	0.08	0.03	0.07	0.15	0.07	0.02	0.05	0.27	0.18	0.07	0.15



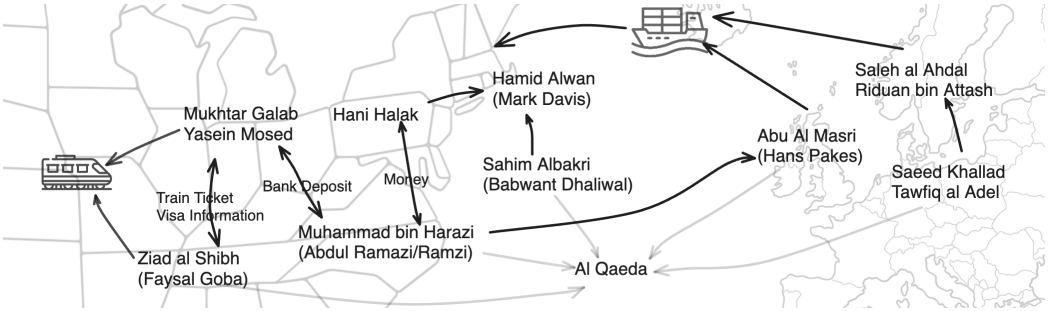
(a) Crescent Dataset



(b) Atlantic Storm Dataset

Fig. 7: Normalized metrics from the narrative for different methods with different temperatures [DET_regular (and/or) DET_person are higher for each metric, across different datasets]

plot the results for combined ROUGE score and METEOR score (appendix). Experiments show that for basic prompts, there are no trends in increasing temperature, for both type of prompts. For the augmented architecture in GPT models, we find two optimum setups for ROUGE and METEOR, i.e., (100 words and 0.7 temperature; 150 words and 0.5 temperature).



(a) Sketch of the solution of Crescent dataset

Abdul Ramazi, owner of Select Gourmet Foods, is involved in money laundering schemes through accounts at Pyramid Bank of Cairo and Central Bank of Dubai. Muhammad Shamzi, arrested in Paris, has connections to individuals using forged passports, including Masood Yaser, Vincent Lozario, and Khalfan Maulid. Ziad al Shibh, also known as Faysal Goba, visited Clark Webster in Richmond, VA. Mukhtar Galab and Yasein Mosed, enrolled at University of Virginia, have expired student visas and have not attended classes. Mukhtar Galab has an account at Virginia National Bank and has been depositing checks from Abdul Ramazi's account. Bagwant Dhaliwal in Queens, NYC, made overseas calls to Hans Pakes in Amsterdam. Hani al Hallak of North Bergen, NJ, deposited checks from Abdul Ramazi's account and was found to have C-4 explosives in his carpet shop. Sahim Albakri, using the alias Bagwant Dhaliwal, and Muhammed bin Harazi, using the alias Abdul Ramazi, have connections to the Taliban. Karm Bensaid, followed by police, was found with binoculars, maps, and flight schedules in a rented truck. Vincent Cortez and Arnaldo Gueterrez, apprehended at the border, had cash and one-way tickets to Houston. Wallace Wilson, an Aryan Nations member, was arrested with ammunition and land mines. 200 pounds of C-4 explosives were reported missing from The Powhatan Company. Jamal Kalifa and Abul Hassan Salman were arrested at an airport in Houston with forged passports. Packages marked as "Home Made Candies" were sent to state government officials in Seattle, potentially containing bombs. Tawfiq al Adel and Saeed Khaliad, arrested in The Netherlands, were connected to Al Qaeda and had radioactive traces in their truck. Saleh al Ahdal and Riduan bin Attash rented a storage unit with radioactive traces and destructive substances. Masood Yaser rented a U-Haul truck with land mines and has an unknown location. The Holland Queen, a container ship bound for Boston, is expected to arrive on April 29. Hans Pakes, working on the Holland Queen, has a history of training with destructive substances. C-4 explosives were found in a carpet shop managed by Hani al Hallak during a fire. Hani al Hallak is currently on vacation in Canada.

(b) DETs augmented output for Crescent Dataset

An terrorist explosives expert Abu al Masri aboard the Holland Queen carries a bomb-triggering device. The cesium carried by Tawfiq al Adel and Saeed Khallad was obtained in Eastern Europe. A quantity of cesium 137 was delivered to Saleh al Ahdal and Riduan bin Attash in Haarlem by Tawfiq Adel and Saeed Khallad. A dirty bomb composed of cesium 137 and TNT was constructed by Saleh al Ahdal and Riduan bin Attash in their storage unit in Haarlem. Saleh al Ahdal and Riduan bin Attash loaded containers carried by the Holland Queen. Muhammed bin Harazi is an Al Qaeda operative. Muhammed bin Harazi met with Abu ai Masri on these visits to Amsterdam. Muhammed bin Harazi directed the activities of Abu al Masri concerning the dirty bomb aboard the Holland Queen. The message 'I will be in my office on 30 April at 9:00 AM. Try to be on time' refers to the time at which terrorist actions will take place. The 200 pounds of C-4 were stolen from the Powhatan Company. Abdulla Ramzi is a double alias for Muhammed bin Harazi [who also uses the alias Abdul Ramazi. The C-4 plastic explosive was stolen by someone associated with Muhammed bin Harazi. Muhammed bin Harazi has made 150 pounds of the stolen C-4 plastic explosive available to Mukhtar Galab, Yasein Mosed, and Ziad al Shibh [alias Faysal Goba]. The three terrorists will each bring 50 pounds of C-4 aboard Train # 19 in their luggage. The three terrorists do not intend to sleep in a single sleeping compartment. A bomb containing 150 pounds of C-4 will be assembled in a sleeping compartment of Train # 19 by the three terrorists. Mukhtar Galab is no longer a student at UVA. Yasein Mosed is no longer a student at UVA. Mukhtar Galab and Yasein Mosed are acting in violation of their student visas. Muhammed bin Harazi is financially supporting Mukhtar Galab and Yasein Mosed in Charlottesville, VA. Mukhtar Galab and Yasein Mosed are involved in terrorist activities. Hani al Halak is a source of explosives for terrorists. The $8500 paid by Harazi to al Halak was a payment for C-4 plastic explosive. Hamid Alwan picked up a quantity of C-4 plastic explosive from Hani al Halak and not a carpet. Hamid Alwan kept the C-4 plastic explosive until he was ready to use it. Hamid Alwan will install C-4 plastic explosive and a timing device in a vending machine [such as a coffee/tea/hot chocolate dispenser] The terrorists brought the C-4 -plastic explosive to their workplace at the Empire State Vending Company [ESVS]. Terrorists will deliver a vending machine, containing a C-4 plastic explosive bomb, to the NYSE on or before 29 April, 2003.

(c) Ground truth narrative for Crescent Dataset [matched: green, and missed:(red: entities, yellow: implicit events)]

Fig. 8: Crescent dataset solution and narratives from LLMs with ground truth

Along with observation from RQ1 & RQ2, it indicates that temperatures in the region of 0.5-1.0 work well. However, there is no improvement in higher temperatures, indicating that the reasoning capability is rather stagnant.

C. Use case studies

Here we will qualitatively showcase the inability for the LLMs to look past the surface-level information. The goal of intelligence analysis is not just to summarize the reports, rather to find out the connections, along with back stories about the adversaries. We showcase the narratives from experiments for Crescent dataset in Fig. 8b and a high level sketch of the solution in Fig. 8a. Crescent dataset plot describes the plans for three synchronized attacks by Al Qaeda operatives, coordinated by Muhammad bin Harazi. Harazi sourced the funds and and organized operatives on three fronts, i) a container ship bound for Boston from Amsterdam, ii) Amtrak Train #19 bound for Atlanta, and iii) the New York Stock Exchange. Ziad al Shibh, Sahim Albakri, Abu al Masri, Saeed Khallad, Tawfiq al Adel are the main individuals, cobwebbed by other operatives on each of three fronts. Reviewing the output against ground truth (Fig. 8c), we find that LLMs lack on two fronts: i) failure to paint out an implicit story about the connections, ii) filtering out relevant and irrelevant documents. We also look at one particular subplot of Crescent dataset. The story is about a ship called "Holland Queen" and how it might have a potentially dangerous cargo onboard. Different versions of our LLMs only described the document-level entities and failed to make a guess that the ship might be carrying dangerous cargo.

V. Recommendations

We answer (**RQ4: Where can LLMs contribute in the 3-step process outlined in section §1**) in the form of recommendations from our findings. From the quantitative and qualitative experiments, it is apparent that the generation of LLMs studied here struggle with in-depth analytical reasoning. Across all datasets, we see trends of good summarization, and less on analytical creativity or speculative/imaginative reasoning (not withstanding LLMs ability to hallucinate information). We briefly describe these points here:

Steering speculative reasoning is hard: LLMs are adept at summarizing documents into groups. However, our repeated attempts with multiple types of prompts and parameter sweeps

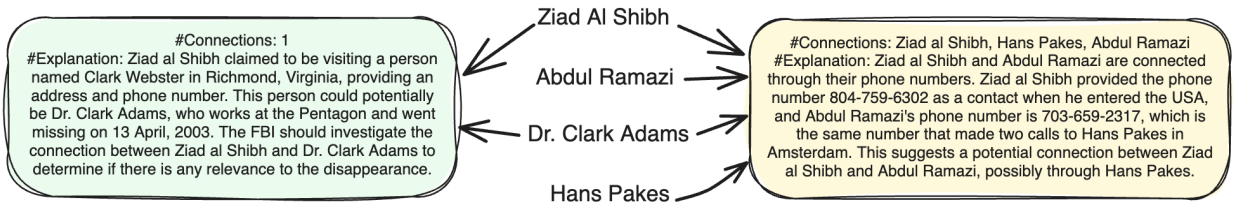


Fig. 9: Small use case for imaginative reasoning: (Left) speculative in nature, combining two people based on the similarity of their name. (Right) failed to invoke the previous speculations about the names of the individuals and strayed away from making any credible speculation at all

failed to invoke the capability required to properly speculate connections. Complementary to this finding, we also experimented with a shorter use case with brief descriptions of four persons as shown in Fig. 9. When asked about the connection between only two persons, LLMs were able to speculate based on the similarity of their names. However, with the addition of two more persons in the mix, LLMs failed to invoke the previous speculation. We also noticed that the position of the target entities mattered. Our finding is consistent with those of recent research, e.g., needle in the haystack [41], lost in the middle [42] studies. Moreover, our findings suggest that, LLMs still struggle on a much smaller scale than previously thought of and this depends on both task complexity (i.e., number of entities in IA) and context length. We plan to address this in future work.

Larger is not necessarily better: Despite GPT-4 being significantly larger, we find no discernible improvement as shown in Table IIIc. The results were consistent across different model families and indicates that the limitation in such reasoning is rather inherent to the LLMs in general. Notably, Gemma-2 scored lower than others, despite being a newer (and supposedly better) model which shows that it is important to establish such real-world and data-centric tasks for improving LLMs in general. As a ray of hope, preliminary tests on OpenAI's o1 model [43] (unavailable during submission) show substantial improvement which can be attributed to the additional chain-of-thought reasoning steps invoked during generation. This reinforces that additional reasoning augmentations can be helpful for IA tasks.

LLMs are good organizers: Even with the lack of in-depth analytical reasoning, LLMs were able to group related entities and events together, specially with smaller datasets. With large datasets, augmentation is needed to orchestrate some of the required grouping because of the context limitation and loss of attention [15].

Orchestrating the evidence is a major challenge: For data-centric analytical tasks like IA, it is hard for an LLM to take everything into context in a single prompt. Augmentation is necessary to organize the evidence. We propose a LLM-driven framework for these tasks. LLMs can also help in the search and retrieval process for such frameworks. Frameworks designed to show the immediate reasoning steps can also be helpful for the analysts to develop mental models.

LLMs can generate good quality reports: Although LLMs throughout our experiments lacked in-depth analytical/imaginative reasoning, they can produce good quality reports with the set of evidences. We suggest the use of LLMs as an interface before and after the data pipeline for such tasks.

TABLE IV: Pitfalls of traditional metrics in intelligence analysis (BERTScore/METEOR/ROUGE-1 scores)

	Crescent	Atlantic Storm	Manpad	Random text
Crescent	X			
Atlantic Storm	0.80/0.22/0.23	X		
Manpad	0.80/0.09/0.23	0.79/0.05/0.12	X	
Random text	0.77/0.06/0.12	0.77/0.03/0.06	0.79/0.12/0.19	X
Atlantic Storm response	0.80/0.20/0.23	0.81/0.21/0.34	0.81/0.19/0.16	0.79/0.15/0.07

Pitfalls of using traditional metrics: We also noticed a significant limitation of traditional metrics, i.e., lexical similarity metrics (ROUGE, METEOR) and contextual embedding based metric (BERTScore) [44]. Due to the common themes, vocabulary, phrases, and content overlap across reports, these metrics exhibit unusually high scores, even across different datasets. This is particularly evident for the Atlantic Storm and Crescent datasets; the most notorious results being that of BERTScore, showing almost equal scores across different datasets. We report the findings in Table IV. In light of these abnormalities, we suggest to use larger foundation models for automatic evaluations [45].

Data condensation is helpful: Turning off data condensation had a noticeably negative impact on experimental results with DETs. This is because DETs rely on concise and precise text for efficient search, retrieval, and other operations. We recommend incorporating data condensation in any framework designed for data-centric applications, as it can significantly improve results.

Classifying relevant/irrelevant documents: One of the challenges for data-centric frameworks such as proposed here is the need for good search and retrieval capabilities; poor performance at this stage percolates downward to lackluster results, a theme well known from retrieval augmented generation (RAG) studies.

Leveraging external knowledge sources: We find some cases where it would be beneficial to have access to an external knowledge base or search engine. One such case arises in the Crescent dataset, where an information dot about reservations on AMTRAK Train #19 was referred. Moreover, the word "Crescent" was also mentioned separately in later reports. Without access to external knowledge or the AMTRAK sched-

ule, an LLM would probably not know that the Train #19 is known as the Crescent Train [46].

Limitations: As our experiments and augmentations are driven by prompt engineering [2], [47], [48] to enable LLMs to carry out text generation tasks, we acknowledge the potential variability in the wording of responses due to changes in prompts. To mitigate this aspect we empirically evaluated a range of prompts to establish a baseline for our studies.

VI. Related Work

From their original roots as text generators [10], [32], LLMs have made their way into a range of downstream applications such as chat agents [16], embodied game players [49], and science solvers [50]. The use of LLMs for IA remains under-explored.

Pirolli and Card [51] were among the first to schematize the process of intelligence analysis using two interacting loops of activities: i) a foraging loop with iterative search and extract of information and ii) a sense-making loop which involves developing a mental model with the information set at hand. Similarly, Klein et al. [52] described the IA process with two components and the interplay between them: i) data as the signals for the events and ii) the frame which explains the events. Sense-making is completed when frames are saturated with the data points. Early approaches for aids were explored in different directions, e.g., model-based [53], multi-agent based [54], entity graph based [22]–[24], and topic modeling based [55]. Various visual analytic systems were developed for entity graph based approaches [25]–[27] and for document level systems [56], [57]. A more recent approach to aid intelligence analysis process comes in the form of using large displays, leveraging virtual reality and immersive displays [9], [58]. All visual analytic approaches however leave the task of sense-making to the analysts.

LLMs have long been augmented by external measures to enhance their intrinsic knowledge [59]–[63]. LLMs have also been designed around external tools to perform real-world tasks [64]–[66] and improve reasoning [67], [68]. There are also tool-based LLMs that can plan and use tools on their own [65], [69]. LLMs are also given external memory module to drive simulation [16]. Our study aims to uncover whether LLMs are suitable as-is or with augmentations to support complex tasks like intelligence analysis.

VII. Conclusion

Strong analytical reasoning and the ability to synthesize information with an imaginative mindset are essential skills for intelligence analysts. LLMs, with their generative strength, should have worked well to aid the analysts. However, as our study shows, effectively connecting the information dots across extensive and complex dossiers appears to be a research frontier. We believe this area holds significant potential for developing better primitives for assimilating knowledge and supporting sensemaking processes for analysts.

References

[1] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[2] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[3] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[4] P. Kumar, "Large language models humanize technology," *arXiv preprint arXiv:2305.05576*, 2023.

[5] J. Kaddour *et al.*, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[6] C. W. Safranek, A. E. Sidamon-Eristoff, A. Gilson, and D. Chartash, "The role of large language models in medical education: applications and implications," p. e50945, 2023.

[7] M. Sharma, A. K. Gogineni, and N. Ramakrishnan, "Neural methods for data-to-text generation," *ACM Transactions on Intelligent Systems and Technology*, 2024.

[8] Y.-a. Kang and J. Stasko, "Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pp. 21–30.

[9] K. Davidson, L. Lisle, K. Whitley, D. A. Bowman, and C. North, "Exploring the evolution of sensemaking strategies in immersive space to think," *IEEE transactions on visualization and computer graphics*, 2022.

[10] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] ——, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[12] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[13] S. Yao *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] M. Besta *et al.*, "Graph of thoughts: Solving elaborate problems with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 682–17 690.

[15] M. Levy, A. Jacoby, and Y. Goldberg, "Same task, more tokens: the impact of input length on the reasoning performance of large language models," *arXiv preprint arXiv:2402.14848*, 2024.

[16] J. S. Park *et al.*, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[17] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022.

[18] J.-T. Huang *et al.*, "Embedding-based retrieval in facebook search," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2553–2561.

[19] H. Su *et al.*, "One embedder, any task: Instruction-finetuned text embeddings," 2023.

[20] A. Q. Jiang *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[21] G. Team *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[22] H. Wu *et al.*, "Where do i start? algorithmic strategies to guide intelligence analysts," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, 2012, pp. 1–8.

[23] E. A. Bier, S. K. Card, and J. W. Bodnar, "Principles and tools for collaborative entity-based intelligence analysis," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 2, pp. 178–191, 2009.

[24] M. S. Hossain, P. Butler, A. P. Boedihardjo, and N. Ramakrishnan, "Storytelling in entity networks to support intelligence analysts," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1375–1383.

[25] E. A. Bier, E. W. Ishak, and E. Chi, "Entity workspace: An evidence file that aids memory, inference, and reading," in *Intelligence and Security Informatics*, S. Mehrotra, D. D. Zeng, H. Chen, B. Thuraisingham, and F.-Y. Wang, Eds., 2006, pp. 466–472.

[26] C. Görg *et al.*, "Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw,"

IEEE transactions on Visualization and Computer Graphics, vol. 19, no. 10, pp. 1646–1663, 2012.
[27] J. Stasko, C. Gorg, Z. Liu, and K. Singhal, "Jigsaw: supporting investigative analysis through interactive visualization," in *2007 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 131–138.
[28] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," 2011.
[29] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996.
[30] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[31] S. Banerjee and A. Lavie, "Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments," *Proceedings of ACL-WMT*, pp. 65–72, 2004.
[32] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[33] L. Zheng *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.
[34] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.
[35] J. Li *et al.*, "Generative judge for evaluating alignment," *arXiv preprint arXiv:2310.05470*, 2023.
[36] H. Huang, Y. Qu, J. Liu, M. Yang, and T. Zhao, "An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers," *arXiv preprint arXiv:2403.02839*, 2024.
[37] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer, "Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings," in *Frontiers in Education*, vol. 8. Frontiers Media SA, 2023, p. 1272229.
[38] Y. Wang *et al.*, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.
[39] S. Diao *et al.*, "Lmflow: An extensible toolkit for finetuning and inference of large foundation models," *arXiv preprint arXiv:2306.12420*, 2023.
[40] A. Alabdulkarim, S. Li, and X. Peng, "Automatic story generation: Challenges and attempts," *arXiv preprint arXiv:2102.12634*, 2021.
[41] gkamradt, "gkamradt/LLMTest_needleinahaystack," Jul. 2024, original-date: 2023-11-11T00:50:02Z. [Online]. Available: https://github.com/gkamradt/LLMTest_NeedleInAHaystack
[42] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.
[43] "Learning to Reason with LLMs." [Online]. Available: https://openai.com/index/learning-to-reason-with-llms/
[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
[45] Y. Liu *et al.*, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.
[46] "Crescent Train New York, Atlanta, New Orleans | Amtrak." [Online]. Available: https://www.amtrak.com/crescent-train
[47] S. Longpre *et al.*, "The flan collection: Designing data and methods for effective instruction tuning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 631–22 648.
[48] S. H. Bach *et al.*, "Promptsource: An integrated development environment and repository for natural language prompts," *arXiv preprint arXiv:2202.01279*, 2022.
[49] G. Wang *et al.*, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv:2305.16291*, 2023.
[50] B. Romera-Paredes *et al.*, "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.
[51] P. Pirolli and S. Card, *The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis*, Jan. 2005.
[52] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, "A Data–Frame Theory of Sensemaking," in *Expertise Out of Context*. Psychology Press, 2007, num Pages: 43.
[53] R. Alonso and H. Li, "Model-guided information discovery for intelligence analysis," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 269–270.
[54] E. Lindahl, S. O'Hara, and Q. Zhu, "A multi-agent system of evidential reasoning for intelligence analyses," in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2007, pp. 1–6.
[55] D. Maiti, M. R. Islam, and N. Ramakrishnan, "Interactive storytelling over document collections," *arXiv preprint arXiv:1602.06566*, 2016.
[56] A. Endert *et al.*, "The human is the loop: new directions for visual analytics," *Journal of intelligent information systems*, vol. 43, pp. 411–435, 2014.
[57] M. S. Hossain, C. Andrews, N. Ramakrishnan, and C. North, "Helping intelligence analysts make connections," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.
[58] L. Lisle *et al.*, "Spaces to think: A comparison of small, large, and immersive displays for the sensemaking process," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2023, pp. 1084–1093.
[59] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
[60] Z. Zhong, T. Lei, and D. Chen, "Training language models with memory augmentation," *arXiv preprint arXiv:2205.12674*, 2022.
[61] Y. Wang, P. Li, M. Sun, and Y. Liu, "Self-knowledge guided retrieval augmentation for large language models," 2023.
[62] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.
[63] K. Shuster *et al.*, "Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion," *arXiv preprint arXiv:2203.13224*, 2022.
[64] P. Lu *et al.*, "Chameleon: Plug-and-play compositional reasoning with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[65] T. Schick *et al.*, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[66] J. Zhang, "Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt," *arXiv preprint arXiv:2304.11116*, 2023.
[67] E. Karpas *et al.*, "Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning," *arXiv preprint arXiv:2205.00445*, 2022.
[68] A. Parisi, Y. Zhao, and N. Fiedel, "Talm: Tool augmented language models," *arXiv preprint arXiv:2205.12255*, 2022.
[69] Y. Qin *et al.*, "Toolllm: Facilitating large language models to master 16000+ real-world apis," *arXiv preprint arXiv:2307.16789*, 2023.

APPENDIX

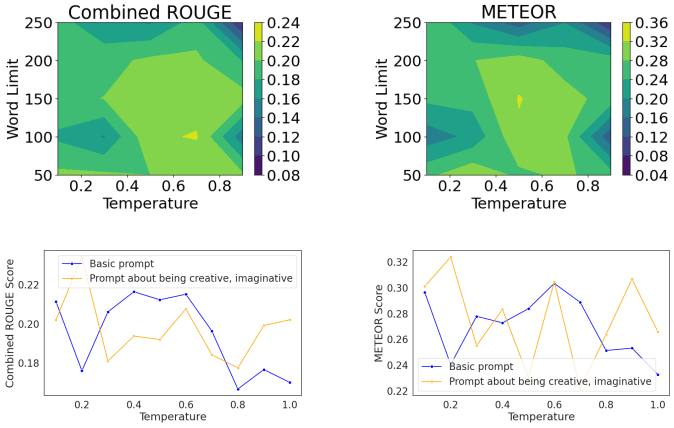


Fig. 10: Effect of temperature, different prompt style, allowable context length on LLM responses