The Irrational LLM: Implementing Cognitive Agents with Weighted Retrieval-Augmented Generation

Dachun Sun, You Lyu, Jinning Li, Xinyi Liu, Denizhan Kara, Christian Lebiere†, Tarek Abdelzaher
Department of Computer Science
University of Illinois Urbana-Champaign, Urbana IL 61801, USA
{dsun18, youlyu2, jinning4, liu323, kara4, zaher}@illinois.edu
† *Department of Psychology*
Carnegie Mellon University, Pittsburgh PA 15223, USA
cl@cmu.edu

***Abstract*—This paper advances research on social networks, extended reality, and the metaverse by bringing together innovations from two different communities – *AI* and *cognitive science* – to develop LLM-based agents with not only fluent responses but also *realistic opinion dynamics* that capture a variety of human biases, imperfections, and general departures from rationality. This avenue of investigation can empower applications from social simulation of human opinions in geopolitical hotspots to realistic non-player character interactions in metaverse games. Recent advances in AI have made remarkable progress toward general intelligence with the introduction of large language models (LLMs). They also enabled grounding LLM responses in specialized information stored externally using retrieval-augmented generation (RAG). In a separate line of research, studies on human cognition have produced cognitive architectures that emulate human departures from rationality, such as biases and imperfections, which are crucial to understanding a wide range of social phenomena and human preferences. A critical mechanism in cognitive architectures is the modulation of retrieval weights from (human) memory; we are biased in what we remember. Combining RAG with cognitive model-inspired computation of information retrieval weights, we develop the *Irrational LLM* – one that weighs information retrieval in RAG systems according to cognitive models, thereby accurately emulating human opinion formation. We implement the novel human cognition-inspired RAG framework (CogRAG) and use it to emulate option developments on different sides of a conflict regarding debated issues. Responses generated by CogRAG (on posts withheld from training data) show close correspondence with real responses posted on social media, suggesting the viability of this approach in approximating biased human opinions. We hope this study paves the way to new directions in AI, social networks, metaverse computing, and human-in-the-loop modeling that better represent diverse human opinions in geopolitical, entertainment, and socio-technical contexts.**

***Index Terms*—Large Language Models, Retrieval-Augmented Generation (RAG), Cognitive Architecture, Bounded Rationality.**

I. Introduction

Despite remarkable advances in AI that led to the development of large language models (LLMs) [1]–[3] with fluent human-like conversational responses [4]–[7], the use of LLMs to emulate human agents (in various simulation contexts) has revealed significant drawbacks and was deemed *perilous* due to their inability to properly reflect artifacts of human psychology and cognition [8], [9]. This paper offers an avenue to help solve this problem by bringing together innovations from two distinct research communities: *AI* and *cognitive science*. Specifically, we interface LLMs borrowed from the AI community (that offer human-like text generation fluency) with cognitive model equations borrowed from cognitive science (that embody the foundations of opinion formation). We interface the two by leveraging a popular mechanism, developed initially to guardrail LLMs from hallucination: Retrieval-Augmented Generation (RAG) [10]–[13]. The RAG framework is a semantically-searchable knowledgebase that stores information to be used by the LLM as external knowledge. By biasing the retrieval probability of different pieces of information stored within the RAG framework (using weights derived from cognitive model equations), we force the LLM to reason using biased premises supplied by the weighted RAG, causing it to adopt biased perspectives on issues in a human-like manner. We call the resulting system, CogRAG.

CogRAG serves as a first step in a longer research agenda for interfacing LLMs, RAG, and cognitive architecture. LLMs (by themselves) typically generate text based on knowledge acquired during training, lacking mechanisms to simulate biased memory recall, an essential trait of human cognition. Moreover, they often struggle with dynamic, real-time contexts, such as political developments or emerging trends. RAG improves the factual accuracy of LLMs by grounding their outputs in retrieved external documents, enabling them to access new information after training. However, conventional RAG assumes a rational retrieval mechanism, returning the most relevant content. In contrast, human cognition tends to favor information that aligns with pre-existing beliefs, has been frequently repeated, or has been recently encountered. This mismatch limits the realism of current LLM-based agents in simulating true human perspectives. CogRAG is a first step to bridge this gap. The current implementation draws from instance-based learning (IBL) theory [14], embodied in a well-studied cognitive architecture, called ACT-R [15], [16]. It models how humans retrieve memories according to recency, frequency, and similarity. By embedding these mechanisms into the RAG pipeline to modulate retrieval in ways that reflect human biases, we demonstrate the ability to produce more realistic agents.

As a case study, we focus on emulating conflicted communities on social media platforms, where posts often elicit starkly

contrasting responses, with celebration from one group and condemnation from another, reflecting ideological divides. As these online interactions increasingly influence offline public behavior and decision-making, a capability for accurately modeling community-specific reactions becomes helpful for many applications, from promoting cross-cultural understanding and improving safety to advancing persuasion and sales. Our evaluation on multiple datasets collected from the $\mathbb{X}$ platform shows that the generated community responses are not only realistic but also more accurately reflect the different sides of the debate and have greater relevance to ongoing events. We hope these results inspire future steps on mitigating the perils of LLM use for social emulation [9].

The rest of this paper is organized as follows. Section II introduces background information on cognitive architecture. Section III presents our problem statement. Section IV details our proposed framework, followed by experimental results in Section V. Section VI reviews related literature, and Section VII concludes the paper and outlines directions for future research.

II. Background

To build cognitively grounded LLM agents that exhibit human-like biases and responses, cognitive models are particularly useful, providing insights into the decision-making processes that drive behavior. A widely used cognitive architecture that implements the basic cognitive mechanisms of information ingestion, recall, and use in various contexts is ACT-R [15]–[17]. It offers a structured framework for modeling human reasoning and decision-making, applicable across various tasks [18], [19]. ACT-R consists of several interacting modules, each modeling different cognitive functions such as working memory, perception, and actions. Declarative knowledge pertains to facts and experience and is the primary focus of this work, as it reflects human memory and significantly influences ideological biases and decision-making.

Declarative memory $\mathbf{M}$ in cognitive models can be viewed as a database storing records of facts and experiences. In ACT-R terminology, each record is referred to as a *chunk*, and each entry or field within a record is known as a *slot*. For instance, people may remember social media posts they have interacted with. Each post is represented as a record (chunk) with slots representing fields such as content, author of the post, and social platform on which the post appeared. To make decisions, chunks are retrieved from declarative memory based on their base activation strength and similarity to the query q in the retrieval buffer. The query is also composed of slots. Let the set of slots in a chunk be $slots$. The activation level A_i of a memory chunk i is determined by:

$$A_i(q) = \ln \sum_{j=1}^{n} (\Delta t_{i,j})^{-\lambda} + K_p \sum_{k \in \text{slots}} \text{sim}_k(v_{i,k}, q_k) + \epsilon_i, \quad (1)$$

The first term reflects the power law of reinforcement and forgetting, where $\Delta t_{i,j}$ is the time elapsed since the j-th activation of chunk i, and λ is the decay factor. The second term accounts for the *partial matching* process, which considers the similarity between the query slot values, $q_k, k \in slots$, in the retrieval buffer, and chunk slot values $v_{\cdot,k}$ in memory, scaled by a mismatch penalty factor K_p. The similarity measure can be defined differently depending on the type of slot value. The general assumption is that the range of the similarity function is $(-\infty, 0]$, so that an exact match results in a similarity value of zero, and mismatches between the query and chunks in the memory decrease the activation level, aligning with cognitive insights. The selection of which slot sets to match depends on the specific problem, and we will present our design in the following section. Lastly, ϵ_i introduces stochasticity in retrieval and is a random value drawn from a logistic distribution with a mean of zero. The probability P_i of retrieving a chunk is based on the activation strength:

$$P_i(q) = \frac{\mathrm{e}^{A_i(q)/T}}{\sum_j \mathrm{e}^{A_j(q)/T}}, \quad (2)$$

where T is the temperature parameter for the softmax equation. We refer to it as *activation probability* later in this work to avoid confusion with the term "retrieval" in RAG.

In social applications, human opinions and preferences often derive from experiential memory, providing an opportunity to leverage instance-based learning (IBL) theory. According to IBL theory [20], [21], decisions or preferences are formed by drawing generalizations from prior experiences, or instances, that resemble the current context. As individuals engage with their environment, they accumulate these instances in memory. When faced with a new situation, such as reading a new post on social media, the system retrieves expectations for each possible action by evaluating the similarity of the current context to past experiences, while also considering the recency and frequency. Newly experienced memories are then also stored in memory, further influencing subsequent decisions.

IBL theory is a domain-general, memory-based theory of experiential learning, and decision-making emerges organically through experiential interaction. Formally, IBL leverages ACT-R's blending mechanism [22] to aggregate multiple memory chunks, creating a blended chunk that interpolates among similar past cases. The blended value $\overline{V_k}$ for slot k represents the consensus stance that best fits past experiences, weighted by their activation probabilities, which is determined by recency, frequency, and situational similarity:

$$\overline{V_k}(q) = \arg\min_V \sum_{i \in \mathbf{M}} P_i(q) \left(1 - \text{sim}_k(V, v_{i,k})\right)^2 \quad (3)$$

where $v_{i,k}$, sim_k, and P_i are defined in Equations 1 and 2.

By leveraging ACT-R and IBL, CogRAG simulates biased memory recall consistent with how humans selectively remember their past experiences and interactions (e.g., selectively remember ideologically aligned content, while deemphasizing other content). Through repeated interaction, each agent develops a memory profile characteristic of that agent's specific past interactions and reflects their biases. This allows our cognitively informed RAG pipeline to retrieve and rank information not just by relevance, but by the way it resonates with the agent's biases, providing a principled foundation for modeling "irrational" LLMs.

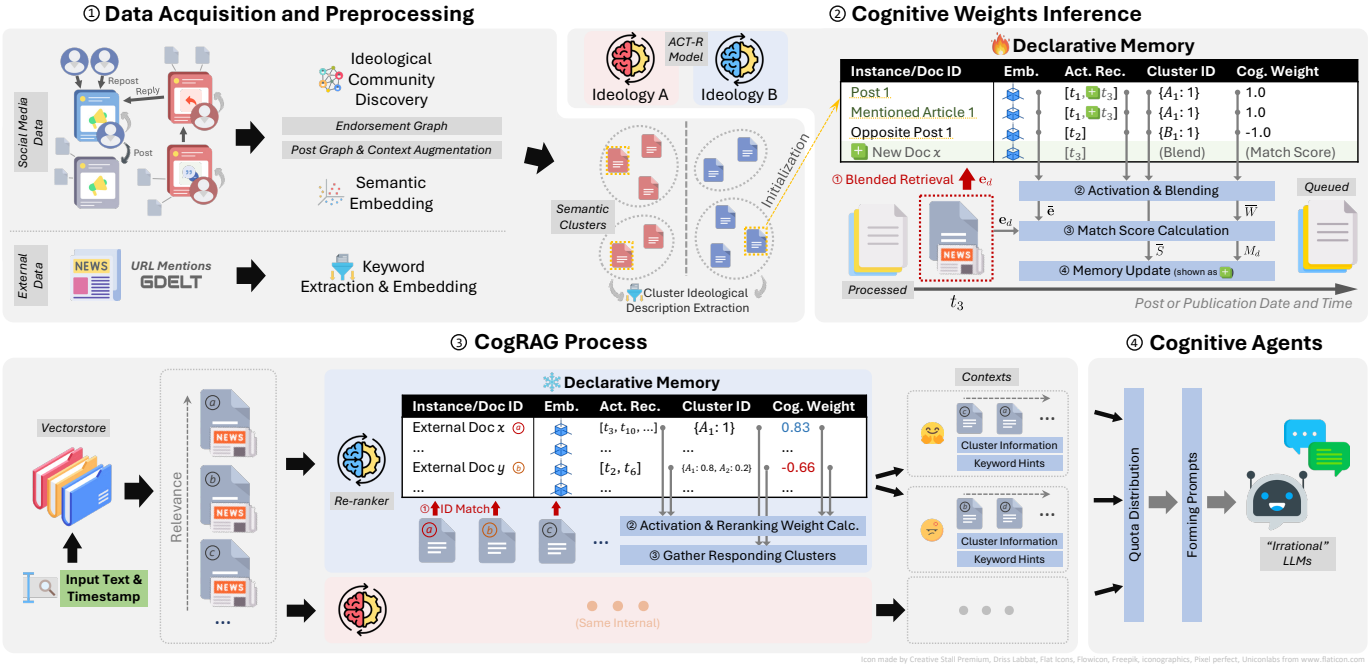

Fig. 1: Framework architecture of CogRAG.

III. Problem Definition

We formally define the task of implementing cognitive agents in a retrieval-augmented setting. Let p_s denote an input social media post to which we want the agents to generate responses, and let $\mathcal{G} = (\mathcal{D}_p \cup \mathcal{U}, E)$ be the social media dataset on a topic, where $\mathcal{D}_p$ are unique posts on the platform and $\mathcal{U}$ is the set of authors. E contains different types of edges, including authorship, endorsing, and replying relations. Also, let $\mathcal{D}_n$ be a continuously updating external news article dataset on the same topic. Suppose there is a set of communities embodying several ideologies, perspectives, or community identities (e.g., liberal versus conservative in a political context or pro- versus anti-positions on a particular issue). The task is to generate a set of realistic responses to input p_s, reflecting the viewpoints and typical reactions of each community. The symbols used in this work are summarized in Table I.

IV. Methodology

We introduce CogRAG, a Cognition-Inspired Retrieval-Augmented Generation framework designed to enable "irrational" LLMs and to implement cognitive agents. Figure 1 illustrates an overview of the proposed framework. Our approach instantiates specialized ACT-R models to learn different ideological preferences, allowing for the inference of cognitive weights on documents that modulate the RAG system's outputs separately for each community. Combining with sampled posts from each community as style references, keyword hints extracted from documents, and an ideologically re-ranked context, we create cognitive agents capable of generating realistic and ideologically diverse responses.

TABLE I: Symbols used and their descriptions.

Symbol	Description
$\mathcal{G}$	Social media dataset on a topic
$\mathcal{D}_p$, $\mathcal{D}_n$	Set of collected posts and news articles
$p.$, $d.$	An augmented social media post or a news article
$\mathbf{M}$	Declarative memory of the ACT-R model with all chunks
$v_{i,k}$	Slot k's value of chunk i
sim_k	Similarity measure for slot k
A_i	Activation level of chunk i
P_i	Activation probability of chunk i
$\overline{V_k}$	Blended value for slot k
H_i, Q_i	Activation records and corresponding popularity of chunk i
C_k, $c_{k,j}$	kth ideological community and jth semantic cluster within
$D_k^{\pm}$	Retrieved context by CogRAG
S_k	Semantic clusters in C_k interested in the retrieved context
θ_A, θ_L, θ_H	Thresholds for activation probability and matching score

A. Instantiating the Cognitive Models

To implement cognitive agents that emulate different ideological perspectives observed in social media communities, we first preprocess the data to prepare an ACT-R-based IBL model for each discovered ideological community by inserting representative posts as initial instances into the declarative memory, thereby priming it with a biased starting point.

1) **Preprocessing:** Many posts in $\mathcal{D}_p$ are part of a cascade and often cannot be fully understood without the conversational context. Therefore, we trace the engagement path along the reply chain in $\mathcal{G}$ and augment each post with the context for its response. Furthermore, given a semantic embedding model $\mathcal{E}$: doc $\to \mathbb{R}^d$, we can embed the augmented posts, as well as external new articles in $\mathcal{D}_n$. To discover ideological communities, we extract an endorsement subgraph from $\mathcal{G}$

containing *who-endorse-what* information, and segment the posts from $\mathcal{D}_p$ into K (usually 2) ideological communities, denoted $C_1, \ldots, C_K$, using a ideological embedding model [23], [24]. Within each ideological cluster C_k, we apply semantic clustering to further divide the posts into more distinct clusters based on semantic embeddings of augmented posts and articles, $C_k = \bigcup_j c_{k,j}$, which encapsulates more consistent themes and viewpoints and facilitates more accurate responses later.

2) **Memory Layout and Initialization:** We define each chunk i in the declarative memory $\mathbf{M}$ of our ACT-R model to include a document identifier, a cognitive weight $v_{i,W}$, a set of soft semantic cluster assignments $v_{i,S}$, an embedding vector $v_{i,\mathbf{e}}$. We also record the chunk activation history H_i and popularity Q_i of the document, defined as the number of engagements up to each activation time. The cluster assignment $(v_{i,S})_j$ is defined as the probability that chunk i belongs to the cluster $c_{i,j}$. For each ideological community C_k, an ACT-R model is instantiated. We select posts that clearly exemplify their beliefs based on ideological embedding, giving them initial cluster assignments according to the previous step and cognitive weights of $+1.0$ to represent ideological alignment. Articles explicitly referenced in these representative posts are also initialized similarly. Additionally, we insert a comparable number of representative posts from the other ideological communities and assign them initial cognitive weights of -1.0, reflecting disagreement or dispute. Each initial memory chunk is assigned an initial activation timestamp equal to its post date and time, preparing the cognitive models to emulate realistic memory decay and reinforcement behaviors.

B. Inferencing More Cognitive Weights

Following initialization, cognitive models representing different ideological communities update their memory dynamically to simulate human-like experiential learning and memory processes by chronologically processing the remaining posts and all external news articles. Inferencing cognitive weights and other slot values for these documents occurs sequentially, simulating real-time attitude changes and memory updates similar to a human reading them one by one.

1) **Blended Retrieval:** For a new document d posted at time t_d, we perform a blended retrieval to calculate three blended values: (1) *cognitive weight* $\overline{V_W}$: the ideological weight representing the document's perceived stance and relevance to community beliefs; (2) *cluster assignment* $\overline{V_S}$: the ideological/semantic clusters that are likely to respond (excluding those with negative cognitive weights); (3) *memory embedding* $\overline{V_\mathbf{e}}$: the blended embedding representing retrieved chunks. We first calculate the activation A_i of chunk i to the document by specializing Equation 1:

$$A_i(d) = \ln \sum_{t_j \in H_{i,<t_d}} \left(\frac{t_d - t_j}{\log_2(1+Q_{i,t_j})} \right)^{-\lambda} + K_p \left[\text{cosine}(v_{i,\mathbf{e}}, d_\mathbf{e}) - 1 \right] \tag{4}$$

where Q_{i,t_j} is the popularity of chunk i up to t_j. We perform partial matching using the shifted cosine similarity on the semantic embedding slot, and $d_\mathbf{e}$ is the embedding of document d. To accelerate the calculation, we approximate Equation 3 for blending as a weighted sum of slot values $v_{i,\cdot}$ based on their activation probabilities P_i, defined in Equation 2:

$$\overline{V_W}(d) \approx \frac{1}{N_A} \sum_{i \in \mathbf{M}, A_i(d) \geq \theta_A} P_i(d) v_{i,W} \tag{5}$$

$$\overline{V_S}(d) \approx \frac{1}{N_A} \sum_{i \in \mathbf{M}, A_i(d) \geq \theta_A} P_i(d) v_{i,S} \tag{6}$$

$$\overline{V_\mathbf{e}}(d) \approx \frac{1}{N_A} \sum_{i \in \mathbf{M}, A_i(d) \geq \theta_A} P_i(d) v_{i,\mathbf{e}} \tag{7}$$

where $N_A = |\{i \in \mathbf{M}, A_i(d) \geq \theta_A\}|$ is the number of activated chunks, and θ_A is the activation threshold. A chunk is not retrieved if its activation is below θ_A, as its contributions to blended values are minimal. This also helps to avoid cluttering the activation history with low-quality records.

Lastly, we calculate a *match score*:

$$W_d = \overline{V_W} \cdot \text{cosine}(\overline{V_\mathbf{e}}, d_\mathbf{e}), \tag{8}$$

which reflects insights from both cognitive science and the RAG paradigm. It is influenced partly by blended cognitive weights that reflect cognitive biases, and partly by the regular similarity between embeddings, typical in RAG systems.

2) **Memory Updates:** If the magnitude of W_d is greater than or equal to the matching threshold θ_L, the experience of processing document d will be stored in memory as a chunk with the embedding $d_\mathbf{e}$, cluster assignment $\overline{V_S}$, and cognitive weight $\widetilde{W}_d$, defined as:

$$\widetilde{W}_d = \begin{cases} W_d & |W_d| < \theta_H \\ \text{sgn}(W_d) & |W_d| \geq \theta_H \end{cases} \tag{9}$$

where θ_H is the threshold for a high match, and $\text{sgn}(\cdot)$ is the sign function. The design addresses the issue of cognitive weights gradually vanishing to zero due to the scaling of cosine similarity. Lastly, we record activations at time t_d for the retrieved chunks, a crucial step in ACT-R models that simulate reinforcement and forgetting. As more external documents are collected, this formulation will facilitate further online updates.

C. CogRAG on Knowledgebase

After presenting all documents chronologically to each ACT-R model, which can be viewed as the meta-persona of each ideological community, we have coded the ideological biases into corresponding cognitive weights and activation history in the model's declarative memory. We then use them as re-rankers to elicit their preferences according to their communities and thus retrieve biased contexts.

1) **Unbiased Candidate Retrieval:** Given an input query post p_s with the associated time t_s, we first perform sparse retrieval using SPLADE [25] on the external document set $\mathcal{D}_n$, obtaining a large enough initial candidate set D_{cand}. Sparse retrieval is term-based and requires lower computational demand, making it particularly helpful when retrieving many longer documents from short-form posts. SPLADE further implements query expansion, improving retrieval relevancy over pure term-based approaches.

2) **Cognitive Re-ranking:** For each cognitive model representing an ideological community C_k, we attempt to match candidates in D_{cand} by identifier to chunks in the memory. Those not present receive zero activation, effectively discarding irrelevant contexts. Matched candidates have their activation probability computed with respect to t_s, incorporating recency and frequency cues. We collect the cluster assignments from them into S_k too, indicating which semantic groups are most interested in the candidates. Matched candidates are split according to the sign of their cognitive weights, $D_k^{\pm}$. Cognitive agents will use one context to support the response, while the other will be the target of refutation. The final re-ranking within $D_k^{\pm}$ employs the product of activation probability and the absolute value of cognitive weights, yielding ideologically important context sets with related semantic clusters $\{D_k^+, D_k^-, S_k\}_{k=1}^K$.

D. The Irrational LLM: Cognitive Agents

We build cognitive agents on top of the CogRAG, combining them with helpful prompting components that explicitly hint at ideological biases and important keywords.

1) **Generation Quota Distribution:** We heuristically allocate a fixed generation quota to each ideological community context, proportional to some function of $|D_{\cdot}^{\pm}|$, and they are further allocated to each semantic cluster, proportional to some function of $\overline{S}_{\cdot}$. To ensure balanced representations even for less active communities, we applied the $\ln(\cdot)$ function to the context set sizes and a linear function to the cluster assignments.

2) **Prompting Components:** We select the top-ranked contextual information from cognitive retrieval and construct a structured prompt for each allocated community and semantic cluster. When generating responses for semantic clusters, we include a few popular posts as writing style references and an (LLM) summarized ideological description of that cluster. To help agents maintain specificity during response generation, we prompted an LLM to extract keyphrases from external articles and applied TF-IDF to identify the most important ones. The above LLM invocations are additional parts of data preprocessing. The full prompt is shown in Figure 2.

```
[SYSTEM]: You will write a reply/response post to a new user-provided post as
    a member of a community. You will also be provided with post examples in
    this community as references.
You know world news and geopolitical matters and like to engage with online
    communications with people from your community/group. You can read
    multiple languages, but you *ALWAYS* write posts in English.
Your writing style SHOULD BE SIMILAR to a Twitter user.

[USER]: The following are examples of how other users in this community post
    original or reply tweets, and you align in general with their standings.
<TWEET_EXAMPLES>"""{Example 1...m}"""</TWEET_EXAMPLES>

[The community is {Summarized ideologies}, described as {descriptions}.]

The following snippets from new articles are the latest information for you
    to know. People in this community like[/dislike] these. You can reference
    them to express yourself in your response.
<NEWS_ARTICLE_SNIPPETS>"""{Snippets 1...n}"""</NEWS_ARTICLE_SNIPPETS>

Try to use the following keywords, but choose the most relevant ones to your
    response: <KEYWORDS>{Keywords}</KEYWORDS>

You read the following post and want to respond to it:
<CONTEXT>{New post content}</CONTEXT>
As a member of this community, write your response:
```

Fig. 2: LLM prompt used by CogRAG with external documents, post samples, ideology description, and keyword hints.

TABLE II: Statistics of social post graph and endorsement graphs of the datasets.

Datasets	Post Graph		Endorsement Graph		
	#Docs	#Edges	#Assertions	#Users	#Edges
COVID	120,743	63,274	48,804	151,857	181,930
RUS	119,231	69,108	93,398	467,363	1,586,572
UK	283,625	96,019	207,696	299,211	973,984
IS-PS	56,016	27,680	29,153	465,507	1,102,255
EDCA	41,670	20,883	26,926	204,556	364,258

V. Experiments

A. Experimental Setup

In this section, we evaluate the performance of the CogRAG framework on five datasets from the $\mathbb{X}$ platform. To evaluate the predictions, we set aside 15 to 30 selected testing posts from each dataset, each with a sufficient number of responses (30-60). CogRAG outperforms other baselines in most settings using various embedding models and LLMs. An example will illustrate the internals of the cognitive agents as well as the responses they generate. Additionally, we demonstrate through an ablation study that all components of the agent's prompt design contribute to its overall performance.

1) Datasets: Below are descriptions of the collected datasets:

- **COVID**: Collected from January to September 2020, during the onset of the COVID pandemic. This dataset is sampled from a public dataset [26]. There are pro- and anti-COVID policy communities.
- **Russophobia (RUS)**: Collected from January to December 2022. The term "Russophobia" is viewed by some as a propaganda term promoted by the Russian government. This dataset also incorporates samples from the public dataset [27]. There are pro- and anti-Russia communities.
- **UK Prime Minister Liz Truss (UK)**: Collected from June to November 2022. This dataset contains posts on the former UK Prime Minister before her election and after her resignation. There are pro- and anti-Truss communities.
- **Israel-Hamas Conflict (IS-PL)**: Collected from October 2023 to January 2024 using keyword filters related to the conflict, such as names of locations and parties involved in the conflict. There are pro- and anti-Israel communities.
- **Enhanced Defense Cooperation Agreement (EDCA)**: Collected from January to June 2023, focusing on EDCA and geopolitical interactions between the Philippines and the US. Two ideological communities are pro-alliance (EDCA supporters, mostly anti-China) and anti-alliance, with the latter further subdivided into two groups: pro-territorial sovereignty and anti-regional escalation.

To use the public datasets, we sampled 100,000 posts within the specified date ranges due to budget constraints. We also use sampled posts as seeds and collect their parent and up to 200 child posts to create a more comprehensive subset. The summary of the dataset statistics is presented in Table II. We also collected 7,366 news articles from mainstream news outlets, such as CNN and Reuters, as external knowledge from mentioned links in social media posts and searching through

TABLE III: Evaluation results of CogRAG against three other baselines with different embedding models for each scenario. LLM is Llama3.3-70B, and we report averages for 30 test cases in each scenario. Bold entries are the best-performing. CogRAG is underlined if it is the second-best-performing. Arrows after metric names indicate whether higher or lower is better.

Embedding	Methods	Emotion JSD ↓					Cluster Matching (%) ↑					ROUGE-L ↑				
		COVID	RUS	UK	IS-PS	EDCA	COV (87)	RUS (72)	UK (60)	IS-PS (75)	EDCA (62)	COVID	RUS	UK	IS-PS	EDCA
VoyageAI	Direct	0.357	0.453	0.437	0.351	0.409	61.67	47.41	42.00	42.92	45.24	0.243	0.341	0.346	0.256	0.235
	Fewshot	0.287	0.315	0.285	0.274	0.326	57.78	55.19	45.67	38.10	45.56	0.292	0.339	0.336	0.261	0.291
	SCRAG	0.265	0.223	0.223	0.230	0.281	58.33	55.56	48.67	47.14	47.50	0.318	0.362	0.350	0.281	0.310
	CogRAG	**0.258**	**0.216**	0.218	**0.214**	**0.232**	**63.75**	**61.11**	**60.83**	**48.89**	**51.43**	**0.331**	**0.382**	**0.376**	**0.297**	**0.328**
OpenAI	Direct	0.356	0.436	0.443	0.364	0.415	60.56	45.56	45.67	38.57	40.00	0.249	0.340	0.342	0.259	0.234
	Fewshot	0.341	0.268	0.281	0.270	0.374	58.89	54.44	46.00	33.33	45.00	0.304	0.331	0.335	0.259	0.296
	SCRAG	0.306	0.263	0.240	0.246	0.267	62.22	55.19	49.33	44.44	49.05	0.314	0.374	0.369	0.270	0.303
	CogRAG	**0.306**	**0.206**	**0.235**	**0.243**	**0.249**	**63.02**	**57.44**	**60.62**	**46.43**	**50.56**	**0.327**	**0.386**	**0.392**	**0.281**	**0.317**
NV-Embed2	Direct	0.375	0.430	0.456	0.383	0.429	63.33	51.48	40.67	34.76	44.79	0.243	0.334	0.343	0.261	0.232
	Fewshot	0.377	0.306	0.270	0.313	0.299	56.67	50.37	45.67	45.00	46.67	0.300	0.333	0.338	0.261	0.291
	SCRAG	0.321	0.273	**0.261**	0.259	0.270	52.67	53.70	52.33	48.33	**51.43**	0.303	0.364	0.361	0.283	0.320
	CogRAG	**0.310**	**0.240**	0.265	**0.231**	**0.244**	**61.25**	**60.83**	**60.21**	**50.00**	49.58	**0.342**	**0.391**	**0.380**	**0.291**	**0.313**

Embedding	Methods	LLM Discrimination Score ↑					Cluster Coverage (%) ↑					Keyword Recall (%) ↑				
		COVID	RUS	UK	IS-PS	EDCA	COVID	RUS	UK	IS-PS	EDCA	COVID	RUS	UK	IS-PS	EDCA
VoyageAI	Direct	7.033	7.822	8.077	7.867	8.037	25.00	58.52	51.33	60.17	61.11	35.71	37.39	32.48	18.95	29.00
	Fewshot	7.050	8.344	8.420	7.940	8.063	26.19	59.26	54.00	65.50	60.42	34.29	34.78	31.62	23.16	32.00
	SCRAG	7.350	**8.552**	**8.967**	8.057	8.227	37.50	**77.78**	66.50	70.56	68.75	38.57	41.74	35.90	25.26	39.00
	CogRAG	**7.356**	8.537	8.876	**8.082**	**8.281**	**38.67**	73.33	**68.89**	**73.12**	**72.22**	**54.29**	**54.78**	**53.85**	**41.05**	**41.00**
OpenAI	Direct	7.039	7.863	7.960	7.970	8.010	25.00	62.04	47.33	56.00	58.33	24.29	36.52	31.62	23.16	32.00
	Fewshot	6.678	8.385	8.587	8.073	8.040	33.33	62.22	63.17	64.67	63.89	34.29	40.87	35.04	25.26	29.00
	SCRAG	7.172	8.430	**8.660**	**8.127**	8.143	37.00	62.96	65.17	62.22	64.29	40.00	41.74	34.19	27.37	34.00
	CogRAG	**7.197**	**8.430**	8.625	8.122	**8.184**	**38.33**	**66.67**	**68.33**	**70.56**	**69.44**	**52.86**	**51.30**	**52.99**	**38.95**	**38.00**
NV-Embed2	Direct	6.720	7.811	7.953	7.823	7.960	27.38	51.85	50.67	48.17	55.56	32.86	37.39	26.50	22.11	28.00
	Fewshot	7.350	8.337	8.677	7.927	8.163	33.00	53.70	65.67	55.67	60.42	31.43	39.13	26.50	20.00	32.00
	SCRAG	**7.707**	8.422	8.693	8.060	**8.310**	**35.00**	64.81	66.50	**70.24**	66.67	32.86	39.13	29.91	26.32	34.00
	CogRAG	7.590	**8.537**	**8.720**	**8.159**	8.223	34.67	**65.62**	**69.63**	66.46	**70.83**	**52.86**	**50.43**	**50.43**	**40.00**	**40.00**

TABLE IV: Evaluation results of CogRAG against fewshot and SCRAG baselines with more LLMs (three models with increasing number of parameters and a commercial model) for each scenario. The embedding model is VoyageAI, and we report averages for 30 test cases in each scenario. Bolding, underlining, and arrows mean the same as in the above table.

LLM	Methods	Emotion JSD ↓					Cluster Matching (%) ↑					ROUGE-L ↑				
		COVID	RUS	UK	IS-PS	EDCA	COV (87)	RUS (72)	UK (60)	IS-PL (75)	EDCA (62)	COVID	RUS	UK	IS-PS	EDCA
Gemma2-9B	Fewshot	0.341	0.247	0.243	0.318	0.273	56.67	59.63	44.00	39.58	47.92	0.334	0.392	0.369	0.307	0.291
	SCRAG	0.290	0.220	0.207	0.226	0.217	60.44	64.81	45.00	**44.33**	48.57	0.351	0.385	0.393	**0.313**	0.341
	CogRAG	**0.277**	**0.154**	**0.185**	**0.212**	**0.197**	**67.71**	**64.88**	**58.33**	42.86	**48.61**	**0.368**	**0.395**	**0.393**	0.306	**0.342**
Qwen2.5-32B	Fewshot	0.380	0.370	0.308	0.359	0.423	62.78	**63.70**	46.00	38.99	41.32	0.267	0.355	0.360	0.207	0.233
	SCRAG	0.362	0.305	0.301	0.293	0.369	67.22	58.15	47.92	42.08	**47.62**	0.331	0.379	**0.395**	0.219	0.301
	CogRAG	**0.344**	**0.246**	**0.299**	**0.251**	**0.323**	**68.33**	57.29	**60.83**	**45.71**	43.75	**0.352**	**0.382**	0.388	**0.222**	**0.311**
Mistral-Large	Fewshot	0.320	0.228	0.236	0.278	0.271	55.56	51.85	48.67	39.58	46.19	0.284	0.365	0.371	0.258	0.249
	SCRAG	0.297	**0.195**	0.201	**0.219**	0.246	61.11	62.59	48.67	44.29	46.67	0.328	0.391	0.364	0.289	0.317
	CogRAG	**0.279**	0.197	**0.199**	0.227	**0.230**	**67.71**	**64.88**	**58.33**	**45.24**	**48.61**	**0.343**	**0.401**	**0.398**	**0.309**	**0.312**
GPT-4o-mini	Fewshot	0.360	0.326	0.287	0.409	0.356	**63.89**	57.41	47.33	28.87	40.97	0.263	0.337	0.329	0.219	0.216
	SCRAG	0.321	0.294	**0.244**	0.300	0.335	60.00	63.33	48.00	39.05	45.42	0.330	0.375	0.379	0.288	0.311
	CogRAG	**0.317**	**0.260**	0.259	**0.244**	**0.316**	62.15	**65.62**	**63.54**	**41.43**	**48.75**	**0.340**	**0.382**	**0.400**	**0.293**	**0.318**

LLM	Methods	LLM Discrimination Score ↑					Cluster Coverage (%) ↑					Keyword Recall (%) ↑				
		COVID	RUS	UK	IS-PS	EDCA	COVID	RUS	UK	IS-PL	EDCA	COVID	RUS	UK	IS-PS	EDCA
Gemma2-9B	Fewshot	6.994	**8.167**	7.947	**8.127**	7.280	28.33	66.67	62.33	60.17	61.67	30.00	37.39	32.48	28.42	35.00
	SCRAG	7.117	7.948	**8.120**	7.900	8.127	40.00	68.52	69.00	67.22	66.67	41.43	39.13	34.19	34.74	38.00
	CogRAG	**7.209**	7.992	8.054	7.969	**8.147**	**46.67**	**69.44**	**70.48**	**72.17**	**68.75**	**54.29**	**55.65**	**54.70**	**45.26**	**50.00**
Qwen2.5-32B	Fewshot	7.044	7.944	8.000	7.870	8.017	27.38	57.41	61.50	57.33	57.41	30.00	34.78	28.21	22.11	29.00
	SCRAG	**7.611**	8.022	**8.057**	8.023	8.130	28.33	68.33	**64.81**	63.17	**66.67**	38.57	40.87	29.91	20.00	29.00
	CogRAG	7.603	**8.043**	8.002	**8.115**	**8.151**	**32.22**	**69.44**	62.67	**66.25**	64.29	**52.86**	**53.04**	**55.56**	**34.74**	**44.00**
Mistral-Large	Fewshot	7.206	8.156	8.433	**8.080**	7.923	26.19	74.07	69.83	60.17	58.33	35.71	30.43	31.62	24.21	33.00
	SCRAG	**7.756**	8.393	8.487	7.827	7.883	34.35	**75.93**	63.17	66.00	66.67	37.14	38.26	35.90	27.37	31.00
	CogRAG	7.610	**8.486**	**8.515**	8.050	**7.961**	**37.78**	73.33	**71.67**	**67.22**	**70.37**	**54.29**	**53.91**	**64.10**	**45.26**	**51.00**
GPT-4o-mini	Fewshot	7.267	7.956	8.407	8.060	7.937	28.33	57.41	58.17	53.96	58.33	30.00	31.30	21.37	16.84	28.00
	SCRAG	7.511	**8.119**	8.447	8.107	8.130	31.67	61.11	60.67	**70.50**	62.96	32.86	39.13	24.79	18.95	28.00
	CogRAG	**7.619**	8.050	**8.481**	**8.150**	**8.154**	**32.22**	**63.10**	**63.33**	69.44	**64.58**	**54.29**	**51.30**	**55.56**	**35.79**	**45.00**

the GDELT project [28]. During the evaluation, we mask the chunks and activations from the future in the cognitive models to avoid data leakage and obtain valid evaluation results.

2) Baselines and Metrics: We adopt similar baselines and metrics as in the SCRAG framework [7], testing CogRAG with multiple embedding models, including VoyageAI (voyage-3-large), OpenAI (text-embedding-3-large), and the open-source NV-Embed-2 model [29]. We also demonstrate its performance with various LLMs, including Gemma2-9B, Qwen2.5-32B, Llama3.3-70B [3], Mistral-Large, and GPT-4o-mini. We compare our approach to the following baseline methods. External knowledge is also included in them to ensure a fair comparison:

- **Direct Prompting**: Prompting the LLM directly and instructing it to predict responses as a social media user.

Fig. 3: Response generation example with sample documents retrieved by CogRAG and cognitive agents' outputs. The generated views expressed above are those of the biased cognitive agents and are not statements of the authors or sponsors.

- **Few-shot Prompting**: Similar to direct prompting, but includes at most five response demonstrations from historical responses under a similar situation.
- **SCRAG**: Construct the prompt by a social computing-inspired RAG, including response examples under similar situations to the query and sparsely retrieved news articles relevant to the query. Ideological communities are formed implicitly using embeddings.

To evaluate the performance of the CogRAG, we adopt the four automatic metrics used in SCRAG and add two metrics that focus on the mention of important keywords and entities to make sure the generated responses are specific and valuable:

- **Emotion JSD & LLM Discrimination Score**: The first metric extracts emotion contents [30] from the generated and real responses and compares the distribution using Jensen-Shannon Divergence (JSD) (bounded by 1). The second prompts the LLM with several real responses and asks it to rate the likelihood of each generated response being real on a scale from 1 to 10.
- **Cluster Matching/Coverage Ratio**: Real responses are clustered based on their embedding vectors, segmenting them as expressing generally different opinions. Generated responses are embedded into the same space, and two ratios are calculated: the percentage of them belonging to one of the clusters and the percentage of clusters covered by at least one predicted response. A high matching ratio indicates relevance and alignment, while a high coverage ratio indicates greater diversity.
- **ROUGE-L**: Measures the longest common subsequence overlap between generated and real responses, reflecting overall content alignment (range between 0 and 1). As this is not really compatible with the generative model evaluation, we calculate this score at the dataset level by joining responses and calculating it using the longer texts.
- **Keyword Recall**: We identify 80 to 120 important phrases and named entities in real responses. This keyword recall reflects how well the model included the specific details or ideological catchphrases that the human used.

B. Evaluation Results

Tables III and IV present the performance of CogRAG across a range of embedding models and LLMs, each generating 30 responses per test case. Overall, CogRAG achieves an average improvement of 7.17% in emotion JSD, 7.3% in cluster matching percentage, 3.4% in cluster coverage percentage, 3.5% in ROUGE-L, and a substantial 48.3% in keyword recall compared to the baselines. Performance on the LLM discrimination score remains comparable to SCRAG, reflecting that both methods leverage LLM capabilities. The gains in emotion JSD, cluster matching, and coverage indicate that CogRAG generates responses that more accurately represent distinct ideological groups, suggesting improved global ideological control and response diversity. Second, although ROUGE-L is not an ideal metric for evaluating open-ended generative tasks, CogRAG still outperforms baselines, likely due to the addition of keyword hints that increase lexical overlap with ground truth. Most notably, the large gains in keyword recall highlight the importance of incorporating key terms through prompt engineering explicitly, and relying solely on examples from prior community posts is inadequate for generating ideologically grounded and topically relevant content.

C. "Irrational" LLM Generation Example

We present a specific working example from the EDCA dataset in Figure 3. The clear and diverse ideological contrasts between communities demonstrate the effectiveness of our framework. The bolded text in the retrieved documents has been manually highlighted to correspond with the assigned cognitive weights. The underlined phrases in the generated responses align with the keywords used in the recall metric. It is noteworthy that nearly all news articles are pro-alliance/pro-EDCA, which is likely due to a general pro-US stance. The ACT-R model for anti-alliance communities can recognize them as highly misaligned with negative cognitive weights and explicitly refute their arguments.

D. Ablation Study

We conducted an ablation study to assess the importance of each component in the prompt of the cognitive agent, using

TABLE V: Evaluation results for the ablation study where CogRAG components are taken offline individually. "Kw" means the keyword hints. "Ref" means sample posts from the community. "Desc" means the community's ideology description.

Methods	Emotion JSD ↓					Cluster Matching (%) ↑					ROUGE-L ↑				
	COVID	RUS	UK	IS-PS	EDCA	COV (87)	RUS (72)	UK (60)	IS-PS (75)	EDCA (62)	COVID	RUS	UK	IS-PS	EDCA
CogRAG	**0.258**	**0.216**	**0.218**	**0.214**	**0.232**	**63.75**	**61.11**	**60.83**	**48.89**	**51.43**	**0.331**	**0.382**	**0.376**	**0.297**	**0.328**
w/o Kw	0.288	0.228	0.238	0.216	0.252	60.42	59.52	59.79	46.43	48.33	0.229	0.325	0.337	0.253	0.231
w/o Ref	0.322	0.236	0.273	0.222	0.272	55.90	54.86	58.13	42.36	45.83	0.262	0.325	0.345	0.258	0.243
w/o Desc	0.304	0.235	0.275	0.235	0.275	52.50	54.46	55.62	43.45	42.08	0.255	0.348	0.348	0.258	0.235
Methods	LLM Discrimination Score ↑					Cluster Coverage (%) ↑					Keyword Recall (%) ↑				
	COVID	RUS	UK	IS-PS	EDCA	COVID	RUS	UK	IS-PS	EDCA	COVID	RUS	UK	IS-PS	EDCA
CogRAG	**7.356**	**8.537**	**8.876**	**8.082**	**8.281**	**38.67**	**73.33**	**68.89**	**73.12**	**72.22**	**54.29**	**54.78**	**53.85**	**41.05**	**41.00**
w/o Kw	7.268	8.252	8.312	7.969	8.172	32.22	60.00	66.00	72.08	71.43	41.43	39.13	29.91	21.05	35.00
w/o Ref	7.213	8.238	8.469	7.892	8.149	34.67	62.50	55.93	75.19	62.50	51.43	50.43	52.99	36.84	40.00
w/o Desc	7.233	8.347	8.323	7.975	8.154	37.78	59.52	52.33	62.33	55.56	52.86	54.78	53.85	35.79	35.00

VoyageAI and Llama 3.3. In Table V, we observe that all components contribute meaningfully to performance. Removing the keyword hint results in significant drops in ROUGE-L and keyword recall, confirming that generations become more generic without these cues, often omitting important entities or ideological talking points. Removing community reference examples forces the LLM to rely solely on internal patterns for generating tweet-style outputs. This affects both LLM discrimination and cluster-based metrics, indicating a reduction in stylistic authenticity. Similarly, removing the community ideology description diminishes the model's ability to ground responses in nuanced ideological positions, especially for ideologies that are not well separated by name alone.

VI. Related Work

RAG combines retrieval techniques with LLMs to enhance performance on knowledge-intensive tasks. By combining parametric memory (the knowledge in the model's weights) with non-parametric memory (the retrieved text), RAG can produce more accurate and specific answers. This concept was evolved in [10]–[13], which integrated retrieval deeply into the generation process for natural language processing tasks that demand extensive knowledge. Sparse retrieval techniques such as TF-IDF and BM25 [31] are more cost-effective but rely on term matching. However, it is an advantage for searching longer news articles based on short texts. SPLADE [25] addresses the accuracy of term matching by incorporating query expansion techniques, making it a better option. Generally, these works do not capture biased notions of relevance, and RAG systems typically retrieve static knowledge rather than the dynamically evolving opinions or reactions that social communities produce. By incorporating ACT-R models, we provide each agent with a distinct perspective. While the retrieved content by CogRAG might be objectively suboptimal or one-sided, that is precisely the point to produce the "irrational LLMs."

ACT-R [15]–[17] is a well-established cognitive architecture that models the interplay between memory, decision-making, and learning processes. Instance-Based Learning (IBL) theory [14] is implemented within the ACT-R cognitive architecture to simulate dynamic decision-making (DDM) [20] by retrieving past instances based on similarity to the current situation. Recent work [19] integrated an ACT-R model with LLMs to create Psychologically-Valid Agents (PVAs) for social simulations. Their model utilized ACT-R's memory retrieval, focusing on recency, frequency, and learned utility of past cases to drive agent decisions. Our framework shares the spirit of combining cognitive models with neural generation, but specifically targets the retrieval mechanism in the generation process. Additionally, we explicitly incorporate confirmation bias by adjusting memory activation based on belief alignment. There is also an emerging interest in analyzing LLMs for human-like biases or irrationalities [32]. However, our work is unique in that we intentionally inject human biases into an LLM system to improve the realism of simulations.

Simulating conversations on social media is generally challenging. The initial data-driven approach [33] used large datasets to learn response patterns. TA-Seq2Seq [34] and context-aware prototype editing [35] improved the relevance of the generated texts by integrating topic information and response prototypes. CGRG [36] has introduced a controllable model for grounded response generation, which permits greater precision in managing responses. There is also interest in making generation more persona-specific, as seen in GLBA [37], which incorporates demographic or group identifiers. Although prior studies have examined information retrieval methods, context-sensitive generation, and factual grounding, they have not explored the possibilities presented by LLMs.A recent work, called SCRAG [7], explored social computing-inspired RAG to generate community responses. Still, it relies heavily on response examples, which poses a challenge in data collection, and lacks explicit controllability of community biases.

VII. Conclusion

In this paper, we presented CogRAG, a cognitively inspired RAG framework that integrates insights from ACT-R architecture. Motivated particularly by biases in human memory recall and remembering, CogRAG modulates external knowledge retrieval processes to reflect these biases, thereby creating an "irrational LLM" capable of simulating realistic discourse on social media. Our experiments on five datasets demonstrate that our method consistently outperforms baselines with excellent robustness across various embedding models and LLMs, generating realistic and ideologically diverse outputs. These results show the practical potential of cognition-inspired

retrieval techniques for modeling human biases and social phenomena in LLM-driven systems. We present a paradigm for enhancing them with bounded rationality. Rather than treating irrationality as noise to be removed, we model these deviations to better understand and predict human behavior. This work paves the way for more comprehensive cognitive integrations in LLMs, where modeling the limits of human rationality is key to replicating (and ultimately understanding) the complex behaviors observed in online communities.

Ethical Statement

Our system is intended solely as a research-oriented tool, helping social scientists, policy-makers, and people alike understand ideological divides, detect polarizations, and guide de-escalation strategies. We acknowledge that simulating human-like cognitive biases could be misused to generate one-sided content. We mandate controlled, research-only deployment, clear labeling of all biased outputs, and expert oversight.

Acknowledgments

Research reported in this paper was sponsored in part by NSF CNS 20-38817, DARPA HR0011-24-3-0325 (BRIES), and the Boeing Company. It was also supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[2] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] Llama Team, AI @ Meta, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[4] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao *et al.*, "DialoGPT: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[6] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, 2024.

[7] D. Sun, Y. Lyu, J. Li, Y. Chen, T. Wang, T. Kimura, and T. Abdelzaher, "SCRAG: Social computing-based retrieval augmented generation for community response forecasting in social media environments," *arXiv preprint arXiv:2504.16947*, 2025.

[8] R. Wang, Z. Huang, S. Liu, H. Shao, D. Liu, J. Li *et al.*, "Dydiff-vae: A dynamic variational framework for information diffusion prediction," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 163–172.

[9] S. Abdurahman, M. Atari, F. Karimi-Malekabadi, M. J. Xue, J. Trager, P. S. Park *et al.*, "Perils and opportunities in using large language models in psychological research," *PNAS Nexus*, vol. 3, 2024.

[10] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 3929–3938.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 9459–9474.

[12] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A survey on retrieval-augmented text generation," *arXiv preprint arXiv:2202.01110*, 2022.

[13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[14] P. Langley and H. A. Simon, "The central role of learning in cognition," in *Cognitive skills and their acquisition*, 2013, pp. 361–380.

[15] J. R. Anderson, *The adaptive character of thought*. Psychology Press, 2013.

[16] ——, *The architecture of cognition*. Psychology Press, 2013.

[17] ——, *Rules of the mind*. Psychology Press, 2014.

[18] C. Lebiere, E. A. Cranford, P. Aggarwal, S. Cooney, M. Tambe, and C. Gonzalez, *Cognitive Modeling for Personalized, Adaptive Signaling for Cyber Deception*. Springer International Publishing, 2023, pp. 59–82.

[19] K. Mitsopoulos, R. Bose, B. Mather, A. Bhatia, K. Gluck, B. Dorr, C. Lebiere, and P. Pirolli, "Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences," *Proceedings of the AAAI Symposium Series*, vol. 2, pp. 340–348, 2024.

[20] C. Gonzalez, J. F. Lerch, and C. Lebiere, "Instance-based learning in dynamic decision making," *Cognitive Science*, vol. 27, pp. 591–635, 2003.

[21] C. Gonzalez, *The boundaries of instance-based learning theory for explaining decisions from experience*. Elsevier, 2013, pp. 73–98.

[22] C. Lebiere, "Blending: An ACT-R mechanism for aggregate retrievals," in *Proceedings of the Sixth Annual ACT-R Workshop*, 1999.

[23] D. Sun, C. Yang, J. Li, R. Wang, S. Yao, H. Shao *et al.*, "Computational modeling of hierarchically polarized groups by structured matrix factorization," *Frontiers in Big Data*, vol. 4, 2021.

[24] J. Li, R. Han, C. Sun, D. Sun, R. Wang, J. Zeng *et al.*, "Large language model-guided disentangled belief representation learning on polarized social graphs," in *IEEE ICCCN 2024*, 2024, pp. 1–9.

[25] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse lexical and expansion model for first stage ranking," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2288–2292.

[26] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set," *JMIR Public Health and Surveillance*, vol. 6, 2020.

[27] E. Chen and E. Ferrara, "Tweets in time of conflict: A public dataset tracking the Twitter discourse on the war between Ukraine and Russia," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, pp. 1006–1013, 2023.

[28] "The GDELT project." [Online]. Available: https://www.gdeltproject.org/

[29] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "NV-Embed: Improved techniques for training llms as generalist embedding models," *arXiv preprint arXiv:2405.17428*, 2024.

[30] R. Plutchik, *Measuring Emotions and Their Derivatives*. Elsevier, 1989, pp. 1–35.

[31] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, 2009.

[32] O. Macmillan-Scott and M. Musolesi, "(Ir)rationality and cognitive biases in large language models," *Royal Society Open Science*, vol. 11, p. 240255, 2024.

[33] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 583–593.

[34] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[35] Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou, "Response generation by context-aware prototype editing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7281–7288, 2019.

[36] Z. Wu, M. Galley, C. Brockett, Y. Zhang, X. Gao, C. Quirk *et al.*, "A controllable model of grounded response generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 14 085–14 093, 2021.

[37] J. Wang, X. Wang, F. Li, Z. Xu, Z. Wang, and B. Wang, "Group linguistic bias aware neural response generation," in *Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing*, Taiwan, 2017, pp. 1–10.