

BEHAVIOR RESEARCH METHODS

Do Machines Remember Like We Do? CogBench-RAG: A Cognitive Benchmark for Retrieval-Augmented Generation Systems

Journal:	<i>Behavior Research Methods</i>
Manuscript ID	Draft
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Saxena, Nikhil; Northeastern University - Boston Campus,

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17 **Do Machines Remember Like We Do?**
18

19 **CogBench-RAG: A Cognitive Benchmark for Retrieval-Augmented Generation Systems**
20
21
22
23
24
25
26
27
28
29
30
31
32
33 **Nikhil Saxena**
34
35 Nikhil Saxena <https://orcid.org/0009-0004-1039-0257>
36
37
38 Correspondence concerning this article should be addressed to Nikhil Saxena,
39
40 Northeastern University, Boston, MA 02115. Email: saxena.ni@northeastern.edu
41
42 The author declares no conflicts of interest. This research received no external funding.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Abstract

Retrieval-Augmented Generation (RAG) systems are increasingly framed as "memory" for large language models, yet no framework exists for evaluating whether their retrieval behavior aligns with human memory phenomena. We introduce CogBench-RAG, a benchmark that tests RAG systems against five memory principles: encoding specificity, the fan effect, proactive and retroactive interference, serial position effects, and retrieval-induced forgetting. We evaluate sparse lexical retrieval (BM25) and dense semantic retrieval using a Cognitive Alignment Score. Both architectures exhibit encoding specificity, monotonically decreasing retrieval accuracy with increasing associative fan, and interference with stronger retroactive than proactive effects. Neither shows serial position effects or retrieval-induced forgetting. The architectures diverge on encoding specificity: BM25 exhibits near-absolute context dependence (mismatch MRR = 0.141), while dense retrieval partially bridges contextual gaps (mismatch MRR = 0.410). These findings suggest that competition-based memory phenomena emerge from similarity-based retrieval, while temporal and practice-dependent effects do not. The null serial position result suggests that retrieval systems do not contribute to "lost in the middle" effects observed in language models. CogBench-RAG provides a reusable framework for evaluating the cognitive properties of retrieval architectures.

Keywords: retrieval-augmented generation, cognitive science, human memory, benchmark, encoding specificity, information retrieval

1
2
3 **Do Machines Remember Like We Do? CogBench-RAG: A Cognitive Benchmark for**
4 **Retrieval-Augmented Generation Systems**

5
6
7 The rapid adoption of Retrieval-Augmented Generation (RAG) has established external
8 retrieval as a primary mechanism for providing large language models with access to knowledge
9 beyond their training data (Lewis et al., 2020). RAG systems are routinely described using
10 memory metaphors: vector databases serve as "long-term memory," retrieval functions as
11 "recall," and the interplay between parametric and non-parametric knowledge mirrors
12 distinctions drawn between semantic and episodic memory systems (Tulving, 1972; Gutierrez et
13 al., 2024). These analogies have practical consequences. Recent systems explicitly draw on
14 neuroscience, with HippoRAG modeling the hippocampal indexing theory of Teyler and
15 Discenna (1986) through knowledge graphs and Personalized PageRank (Gutierrez et al., 2024),
16 EM-LLM implementing surprise-based episodic segmentation from event cognition research
17 (Fountas et al., 2025; Zacks et al., 2007), and ARM incorporating Ebbinghaus-style memory
18 decay and consolidation (Bursa, 2026; Ebbinghaus, 1885).

19
20 Despite this convergence between retrieval system design and cognitive science, an open
21 question persists: do RAG systems actually behave like human memory? The existing literature
22 uses cognitive science as design inspiration for building systems, but no work has systematically
23 used cognitive science as an evaluation framework for understanding them. If retrieval systems
24 exhibit human-like memory biases, practitioners need to anticipate and mitigate them. If they do
25 not, the memory metaphors guiding system design may be misleading.

26
27 Human memory research offers over a century of rigorously characterized phenomena,
28 from Ebbinghaus's (1885) foundational work on forgetting curves to contemporary models of
29 retrieval-induced forgetting (Anderson et al., 1994). These phenomena are robust, well-

1
2
3 quantified, and replicated across diverse materials and populations (Baddeley et al., 2015;
4
5 Kahana, 2012). They provide precise behavioral predictions that can be operationalized as
6
7 benchmark tasks. Yet no retrieval benchmark, including BEIR (Thakur et al., 2021), MTEB
8
9 (Muennighoff et al., 2022), RAGAS (Es et al., 2024), or ARES (Saad-Falcon et al., 2024), tests
10
11 whether retrieval systems exhibit these patterns.
12
13

14 We introduce CogBench-RAG, a benchmark suite comprising five modules, each
15
16 grounded in a foundational human memory phenomenon: (a) encoding specificity (Tulving &
17
18 Thomson, 1973), where retrieval success depends on the match between encoding and retrieval
19 contexts; (b) the fan effect (Anderson, 1974), where retrieval accuracy decreases as the number
20
21 of facts associated with a concept increases; (c) proactive and retroactive interference
22
23 (Underwood, 1957; Muller & Pilzecker, 1900); (d) serial position effects (Murdock, 1962;
24
25 Ebbinghaus, 1885); and (e) retrieval-induced forgetting (Anderson et al., 1994).
26
27
28

29 We evaluate BM25 (sparse, lexical; Robertson & Zaragoza, 2009) and dense retrieval
30
31 using sentence-transformers (semantic, embedding-based; Reimers & Gurevych, 2019) and find
32
33 a pattern of selective alignment. Both systems exhibit human-like encoding specificity, fan
34
35 effects, and interference, but neither shows serial position effects or retrieval-induced forgetting.
36
37 The two architectures also diverge in how they exhibit shared phenomena: BM25 shows near-
38
39 absolute encoding specificity where contextual mismatch is equivalent to retrieval failure, while
40
41 dense retrieval partially bridges contextual gaps through semantic similarity.
42
43
44

45 Our contributions are: (a) CogBench-RAG, an open-source benchmark suite that
46
47 operationalizes five human memory phenomena as retrieval evaluation tasks; (b) the Cognitive
48
49 Alignment Score (CAS), a normalized metric for quantifying human-likeness of retrieval
50
51 behavior; (c) a systematic mapping between RAG system behavior and established human
52
53
54
55
56
57
58
59
60

memory phenomena, suggesting selective alignment that varies by both phenomenon and architecture; and (d) evidence that standard retrieval architectures do not contribute to "lost in the middle" effects (Liu et al., 2024) and that encoding specificity varies systematically between lexical and semantic retrieval.

Figure 1: CogBench-RAG Framework

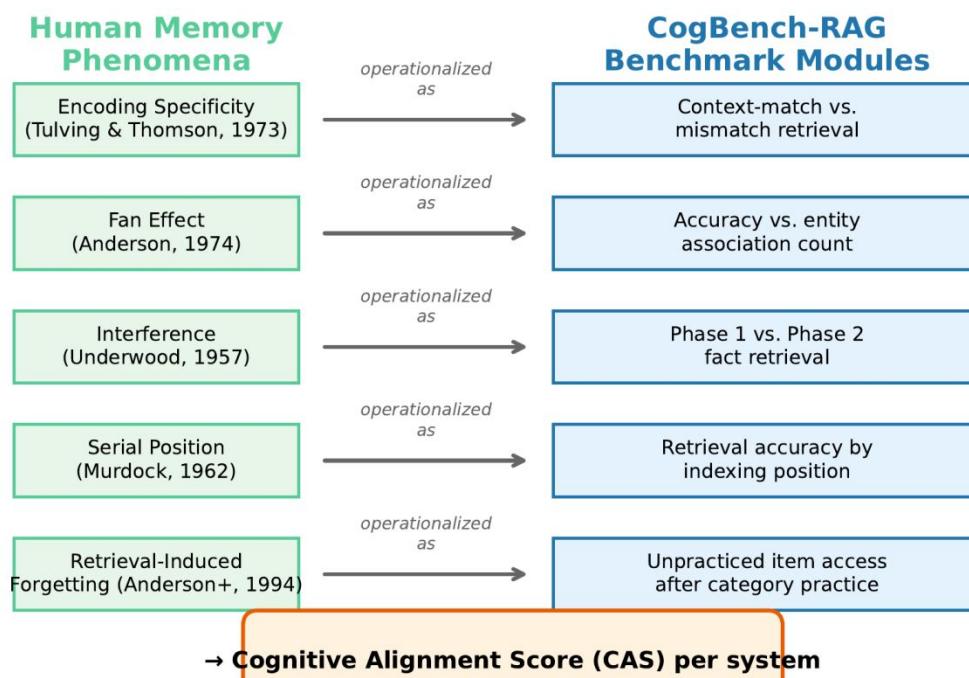


Figure 1. CogBench-RAG framework mapping five human memory phenomena to benchmark modules. Each phenomenon is operationalized as a controlled retrieval task, and a Cognitive Alignment Score (CAS) is computed per system per module.

Related Work

Retrieval-Augmented Generation

RAG was introduced by Lewis et al. (2020), who combined a pre-trained seq2seq model with a dense passage retriever to ground language generation in retrieved evidence. Since then,

the approach has diversified. Dense passage retrieval using learned dual-encoder embeddings (Karpukhin et al., 2020) coexists with sparse lexical methods such as BM25 (Robertson & Zaragoza, 2009), hybrid sparse-dense approaches (Ma et al., 2021), and graph-augmented systems (Edge et al., 2024). Iterative retrieval methods such as FLARE (Jiang et al., 2023) trigger retrieval based on generation-time uncertainty, while Self-RAG (Asai et al., 2024) learns when to retrieve and how to critique retrieved passages.

Recent work has moved toward cognitively-inspired architectures. HippoRAG (Gutierrez et al., 2024) mimics the hippocampal memory indexing theory (Teyler & Discenna, 1986) using knowledge graphs and Personalized PageRank. EM-LLM (Fountas et al., 2025) applies surprise-based event segmentation from cognitive models of episodic memory (Zacks et al., 2007; Radvansky & Zacks, 2014). Bursa (2026) introduces Adaptive RAG Memory (ARM), implementing selective remembrance and decay inspired by memory consolidation (McGaugh, 2000). These systems reflect growing interest in cognitive alignment but evaluate success through task performance metrics rather than behavioral correspondence with human memory.

Retrieval Benchmarks

BEIR (Thakur et al., 2021) established the standard for zero-shot retrieval evaluation across 18 heterogeneous datasets. MTEB (Muennighoff et al., 2022) extended this to embedding evaluation across classification, clustering, reranking, and semantic similarity. RAG-specific evaluations have also expanded: RAGAS (Es et al., 2024) assesses faithfulness and context precision; ARES (Saad-Falcon et al., 2024) provides automated RAG evaluation; and RGB (Chen et al., 2024) benchmarks RAG robustness. All existing benchmarks evaluate task performance rather than behavioral characterization. CogBench-RAG addresses a different

1
2
3 question: whether a system's retrieval behavior exhibits specific patterns predicted by cognitive
4 theory, independent of its performance on any particular downstream task.
5
6

7
8 **Cognitive Evaluation of AI Systems**
9

10 A growing body of work tests whether AI systems replicate human cognitive patterns.
11
12 Hagendorff et al. (2023) found that GPT-3 exhibits human-like intuitive behaviors on the
13 Cognitive Reflection Test. Koo et al. (2024) benchmarked cognitive biases in LLM evaluation
14 outputs, finding 40% of comparisons exhibited biases. Cheung et al. (2025) observed amplified
15 omission bias in LLM moral decision-making. Binz and Schulz (2023) evaluated GPT-3 against
16 human decision heuristics. Suri et al. (2024) investigated anchoring effects in LLMs. Kim et al.
17 (2025) found that reasoning capabilities did not protect against clinical cognitive biases.
18
19

20 This work has focused on language model generation behavior. CogBench-RAG applies
21 the same approach to the retrieval stage, testing whether systems that supply information to
22 LLMs exhibit their own cognitive patterns.
23

24
25 **Memory and Retrieval in Cognitive Science**
26

27 The five phenomena we test represent well-established findings in cognitive science.
28
29 Encoding specificity (Tulving & Thomson, 1973) has been observed across verbal, spatial, and
30 environmental contexts (Godden & Baddeley, 1975; Eich, 1980; Marian & Neisser, 2000). The
31 fan effect (Anderson, 1974) is a core prediction of ACT-R (Anderson et al., 2004; Anderson &
32 Reder, 1999) and has been documented across propositional, spatial, and visual materials
33 (Radvansky et al., 1993). Interference theory has been central to understanding forgetting since
34 Muller and Pilzecker (1900), with the retroactive-exceeds-proactive asymmetry consistently
35 replicated (Wixted, 2004; Kliegl & Bauml, 2021; Postman & Underwood, 1973; McGeoch,
36 1932). Serial position effects (Murdock, 1962) arise from differential rehearsal (Rundus, 1971)
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 and recency of activation in working memory (Glanzer & Cunitz, 1966; Atkinson & Shiffrin,
4
5 1968). Retrieval-induced forgetting (Anderson et al., 1994; Anderson, 2003) has been confirmed
6
7 via meta-analysis across over 200 experiments (Murayama et al., 2014).
8
9

10 **Method**
11

12 **Overview**
13

14 CogBench-RAG comprises five benchmark modules, each operationalizing a human
15 memory phenomenon as a controlled retrieval task. Each module generates a corpus and query
16 set organized into experimental conditions. Retrieval performance is measured using Mean
17 Reciprocal Rank (MRR) and Recall@1, and a Cognitive Alignment Score (CAS) quantifies
18 pattern correspondence with the expected human-like pattern. All experiments use controlled
19 synthetic corpora with fixed random seeds (numpy seed = 42, PYTHONHASHSEED = 0)
20 following established benchmarking methodology (Germain et al., 2020; Luecken & Theis,
21 2019).
22
23

24 **Systems Under Test**
25

26 We evaluate two architecturally distinct retrieval systems. BM25 (Robertson & Zaragoza,
27 2009) is a sparse lexical retrieval method based on the probabilistic relevance framework, with
28 parameters $k_1 = 1.5$ and $b = 0.75$ (Manning et al., 2008). It operates on exact token matches and
29 is invariant to semantic similarity between non-identical terms.
30
31

32 Dense Retrieval (MiniLM) is a semantic retrieval method using the all-MiniLM-L6-v2
33 sentence-transformer model (Wang et al., 2020; Reimers & Gurevych, 2019) to encode
34 documents and queries into 384-dimensional embeddings. Retrieval uses cosine similarity via
35 FAISS (Johnson et al., 2019) exact inner product search on L2-normalized vectors. Both systems
36 use top-k = 10 retrieval.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Module 1: Encoding Specificity

Tulving and Thomson (1973) showed that memory retrieval is most effective when cues present at retrieval match those present during encoding. This principle has been supported across environmental contexts (Godden & Baddeley, 1975), emotional states (Eich, 1980), and linguistic contexts (Marian & Neisser, 2000).

We construct eight topic pairs, each expressing the same core fact in two different domain framings. For example, the effects of rising ocean temperatures on marine ecosystems are expressed in both a marine biology framing (discussing zooxanthellae, thermal stress, and coral bleaching) and an economics framing (discussing seafood industry losses, tourism revenue, and insurance markets). This yields 16 documents. For each document, three query conditions are tested: context match (same domain vocabulary), context mismatch (other domain's vocabulary for the same underlying information), and unrelated (different topic entirely). This yields 48 queries.

Module 2: Fan Effect

Anderson (1974) showed that retrieval accuracy decreases as the number of facts associated with a concept increases, a core mechanism in ACT-R (Anderson et al., 2004; Anderson & Reder, 1999). We generate 80 entities with fan sizes of 1, 2, 5, and 10 (20 entities each), yielding 360 documents. Queries reference an entity and its associated domain keyword, creating genuine retrieval competition.

Module 3: Interference

Interference theory has been central to understanding forgetting since Muller and Pilzecker (1900). The retroactive-exceeds-proactive asymmetry is a consistent finding (Wixted, 2004; Kliegl & Bauml, 2021), and interference is modulated by similarity (McGeoch, 1932;

Osgood, 1949). Six entities each have two documents representing sequential phases, with four control entities appearing in only one phase. Queries use generic shared attributes analogous to the A-B, A-C paired-associate paradigm (Barnes & Underwood, 1959), yielding 32 queries.

Modules 4 and 5: Serial Position and Retrieval-Induced Forgetting

The serial position effect (Murdock, 1962; Glanzer & Cunitz, 1966) produces a U-shaped recall curve with primacy attributed to rehearsal (Rundus, 1971) and recency to working memory activation (Atkinson & Shiffrin, 1968). We construct two 10-document sequences, motivated additionally by the "lost in the middle" finding (Liu et al., 2024). Anderson et al. (1994) showed that practicing retrieval of some category members suppresses access to unpracticed members, attributed to inhibitory control (Anderson, 2003; Murayama et al., 2014). We construct four document categories with practiced and unpracticed items plus a baseline category.

Cognitive Alignment Score

For each module, we compute a CAS normalized to [0, 1], where 1.0 indicates correspondence with the human behavioral pattern and 0.0 indicates no alignment. Encoding specificity CAS weights the match-mismatch MRR gap (60%) and correct condition ordering (40%). Fan effect CAS uses the Spearman correlation between fan size and MRR. Interference CAS combines PI and RI effect magnitudes relative to controls. Serial position CAS detects primacy and recency advantages. RIF CAS measures the baseline-test difference.

Results

1
2
3
4 **Figure 2: Cognitive Alignment Across Systems and Phenomena**
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

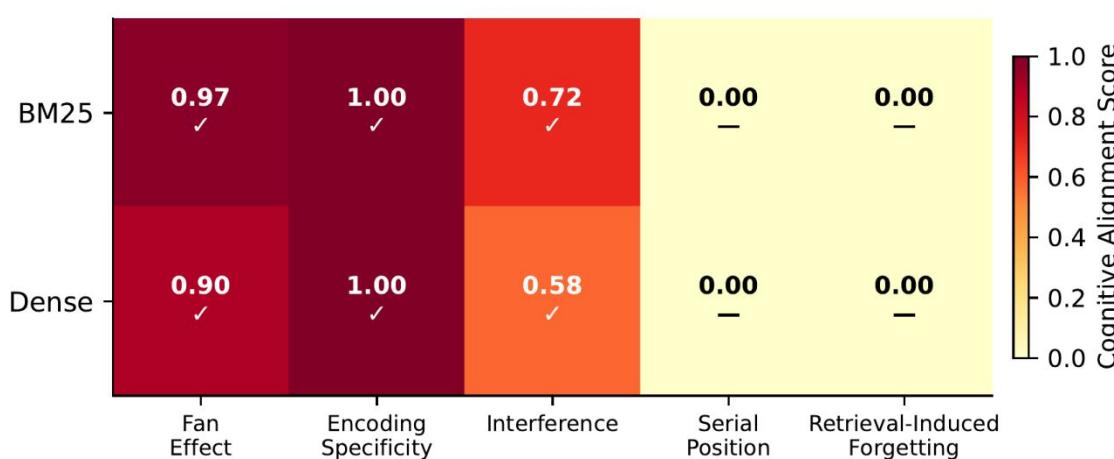


Figure 2. Cognitive Alignment Scores across systems and phenomena. Checkmarks indicate human-like patterns; dashes indicate null results. Both systems achieve high alignment on encoding specificity and the fan effect, moderate alignment on interference, and null alignment on serial position and retrieval-induced forgetting.

Encoding Specificity

Both systems exhibited encoding specificity, but with qualitatively different profiles (Figure 3). BM25 showed near-absolute context dependence: context-matched queries achieved perfect retrieval ($MRR = 1.000$), while context-mismatched queries performed at the level of unrelated queries (mismatch $MRR = 0.141$, unrelated $MRR = 0.140$). For BM25, querying about ocean warming effects using marine biology vocabulary versus economics vocabulary proved no more effective than querying about an entirely unrelated topic.

Dense retrieval also showed encoding specificity but with partial bridging of the contextual gap. Context-matched queries achieved perfect retrieval ($MRR = 1.000$), while context-mismatched queries performed above the unrelated baseline (mismatch $MRR = 0.410$,

unrelated MRR = 0.101). Dense retrieval captures semantic overlap between domain-mismatched descriptions of the same phenomenon that BM25 cannot access.

Both systems achieved CAS = 1.00. The encoding specificity effect was larger for BM25 (match-mismatch delta = 0.859) than for dense retrieval (delta = 0.590), indicating that lexical retrieval is more context-dependent than semantic retrieval. This is consistent with Tulving and Thomson's (1973) observation that retrieval success depends on cue-encoding overlap, with the degree of dependence varying by the nature of the representation. The pattern is also interpretable through the levels-of-processing framework (Craik & Lockhart, 1972): BM25's surface-level processing produces context-bound representations, while dense retrieval's semantic processing produces more transferable traces.

Figure 3: Encoding Specificity Effect by Retrieval Architecture

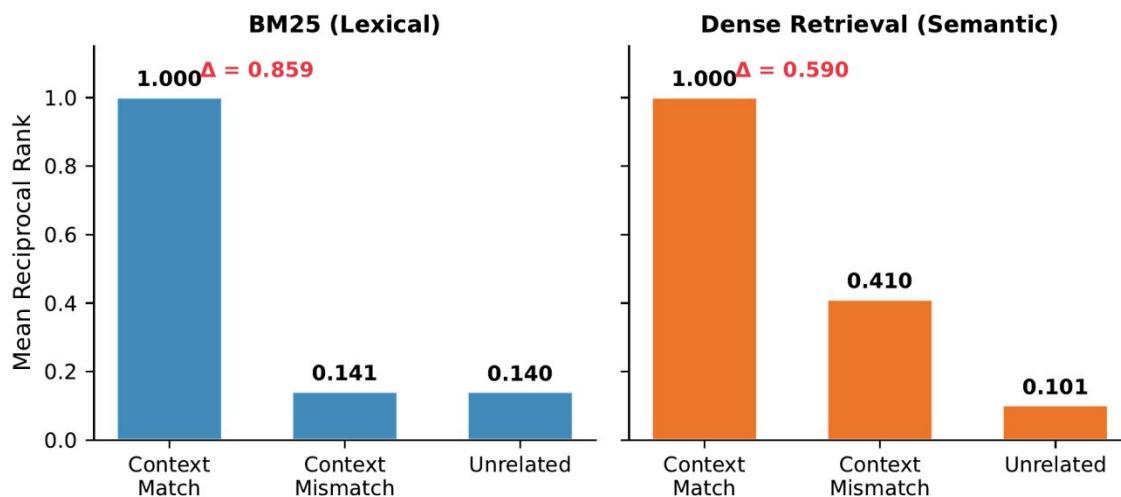


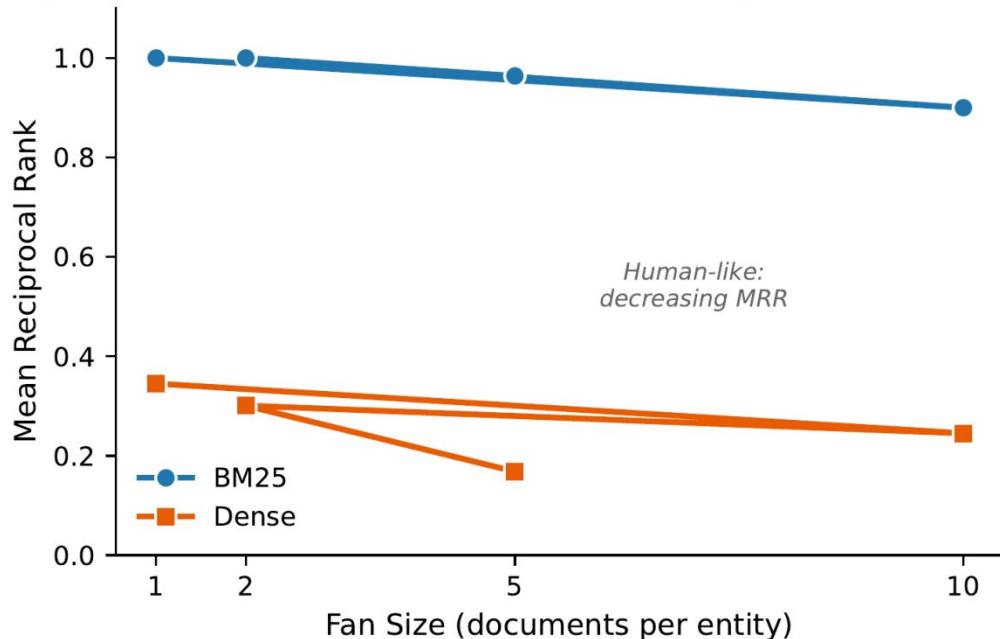
Figure 3. Encoding specificity effect by retrieval architecture. BM25 (left) shows near-absolute context dependence, with context-mismatch MRR (0.141) equivalent to unrelated queries (0.140). Dense retrieval (right) partially bridges the contextual gap (mismatch MRR = 0.410 vs. unrelated = 0.101). Delta values indicate the match-mismatch MRR difference.

1
2
3 **Fan Effect**
4
5

6 Both systems showed decreasing retrieval accuracy with increasing associative fan
7 (Figure 4). BM25 exhibited a monotonic decline from MRR = 1.000 at fan size 1, through MRR
8 = 1.000 at fan size 2 and MRR = 0.964 at fan size 5, to MRR = 0.900 at fan size 10 (Spearman
9
10 $r_s = -0.949$). Dense retrieval showed a comparable pattern: MRR declined from 0.345 at fan
11 size 1, through 0.301 at fan size 2 and 0.168 at fan size 5, to 0.245 at fan size 10 ($r_s = -0.800$).
12
13
14

15 Both correlations indicate a human-like fan effect consistent with the core pattern from
16 Anderson's (1974) original experiments and with ACT-R's prediction that retrieval competition
17 increases with the number of associated items (Anderson et al., 2004). BM25 maintains higher
18 absolute performance across fan sizes, reflecting its lexical precision advantage on documents
19 containing unique domain keywords, rather than a difference in the underlying pattern.
20
21
22
23
24
25
26
27
28

29 **Figure 4: Fan Effect — Retrieval Accuracy vs. Associative Fan**



52 *Figure 4.* Fan effect: retrieval accuracy (MRR) versus associative fan size. Both BM25 and
53
54 dense retrieval show the human-like pattern of decreasing accuracy with increasing fan. The
55
56
57
58
59
60

1
2
3 trend line shows the overall decreasing relationship; BM25 operates at higher absolute MRR due
4
5 to lexical precision on the synthetic corpus.
6
7
8

9 Interference 10

11 Both systems exhibited proactive and retroactive interference, with an asymmetry
12
13 matching the human pattern (Figure 5). Control entities, appearing in only one phase, achieved
14
15 perfect retrieval ($MRR = 1.000$ for both systems).
16
17

18 For BM25, retroactive interference was substantially stronger than proactive interference
19
20 (RI: $MRR = 0.521$; PI: $MRR = 0.833$), representing drops of 0.479 and 0.167 relative to control,
21
22 respectively. For dense retrieval, the same asymmetry was observed (RI: $MRR = 0.612$; PI:
23
24 $MRR = 0.917$), though both effects were smaller than in BM25.
25
26

27 This retroactive-exceeds-proactive asymmetry is one of the most consistent findings in
28
29 interference theory (Underwood, 1957; Postman & Underwood, 1973; Wixted, 2004). Its
30
31 emergence in retrieval systems not designed to exhibit it suggests the asymmetry arises from
32
33 structural properties of competitive retrieval. Phase 2 documents describe more recent events
34
35 with vocabulary reflecting current activities, which tends to overlap more with generic present-
36
37 tense queries than Phase 1's historical descriptions, creating a recency-favoring competition
38
39 dynamic.
40
41

42 BM25 showed stronger interference overall ($CAS = 0.72$) than dense retrieval ($CAS =$
43
44 0.58), consistent with greater susceptibility to competition from lexically overlapping content.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

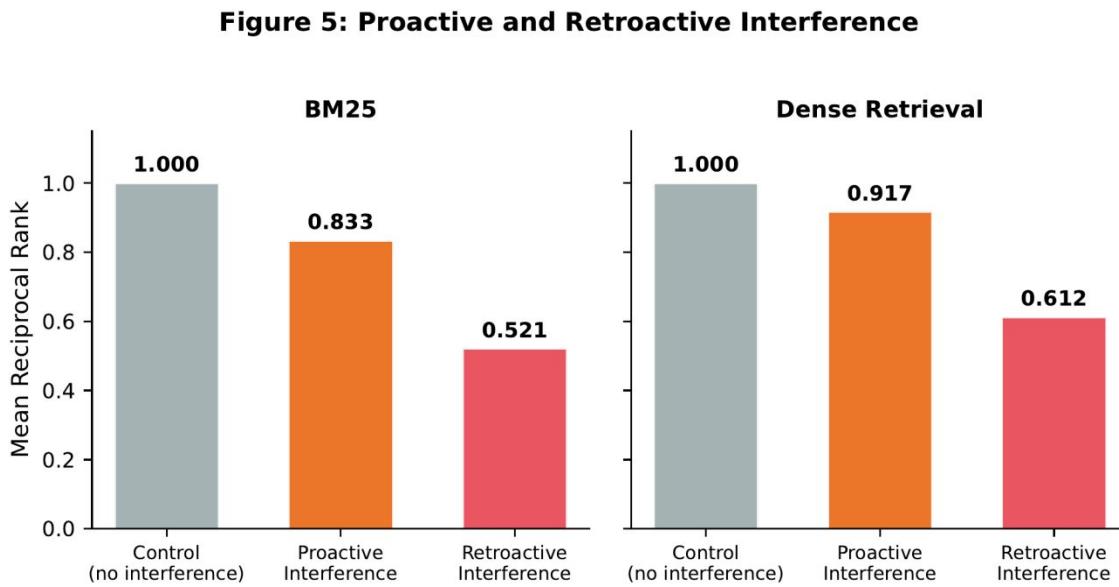


Figure 5. Proactive and retroactive interference effects. Both systems show the human-like pattern of stronger retroactive than proactive interference. Control entities (single phase) achieve perfect retrieval (MRR = 1.000). BM25: PI MRR = 0.833, RI MRR = 0.521. Dense: PI MRR = 0.917, RI MRR = 0.612.

Serial Position

Neither system exhibited serial position effects. Both BM25 and dense retrieval achieved MRR = 1.000 at every sequence position (CAS = 0.00). Both lexical and semantic retrieval architectures are order-invariant: the position at which a document was indexed has no effect on retrieval. BM25's inverted index and FAISS's flat inner product search are both invariant to insertion order by construction.

The result is relevant to the "lost in the middle" phenomenon (Liu et al., 2024), in which LLMs underweight information in the middle of long contexts. Our results indicate that this positional bias originates in the generation stage rather than the retrieval stage. Standard RAG retrieval does not introduce serial position biases before the LLM processes retrieved content.

Mitigation efforts should therefore focus on the generation stage, such as passage reordering or position-aware attention (Peysakhovich & Lerer, 2023), rather than retrieval modification.

Retrieval-Induced Forgetting

Neither system exhibited retrieval-induced forgetting. Unpracticed items in practiced categories were retrieved with identical accuracy to baseline items ($MRR = 1.000$, $CAS = 0.00$). The null result follows from the stateless nature of both systems. BM25 and dense retrieval do not modify their indices based on prior queries, so retrieving one document cannot affect the accessibility of related documents. Standard RAG systems are therefore immune to the inhibitory suppression mechanism proposed by Anderson (2003). However, this immunity may not persist in emerging adaptive architectures with session-dependent caching or reinforcement learning-based re-rankers (Chen et al., 2025).

Summary

Across all five modules, both systems exhibited three human-like patterns (encoding specificity, fan effect, interference) and two null results (serial position, RIF). The overall pattern indicates that competition-based memory phenomena emerge from similarity-based retrieval, while temporal and practice-dependent effects require mechanisms absent in standard architectures.

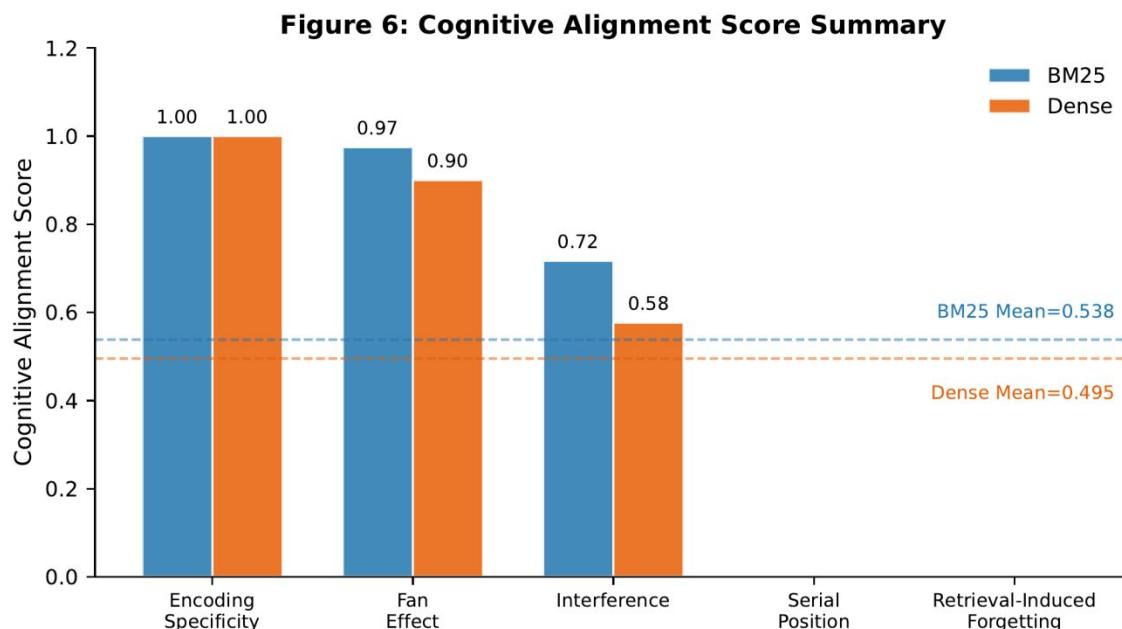


Figure 6. Cognitive Alignment Score summary across all five modules for BM25 and dense retrieval. BM25 mean CAS = 0.538; Dense mean CAS = 0.495. Both systems show high alignment on encoding specificity and the fan effect, moderate alignment on interference, and null alignment on serial position and retrieval-induced forgetting.

Discussion

Selective Cognitive Alignment

The results suggest that RAG retrieval systems exhibit a specific subset of human memory phenomena, namely those arising from competition during retrieval, while being unaffected by phenomena arising from temporal dynamics and practice-dependent plasticity.

Encoding specificity, the fan effect, and interference all emerge from competition among candidate documents. This competition is inherent to similarity-based ranking and produces patterns paralleling human memory retrieval, consistent with the principle that competition is a fundamental constraint on any retrieval system, biological or artificial (Anderson et al., 2004; Watkins & Watkins, 1975).

1
2
3 Serial position effects and retrieval-induced forgetting require temporal or state-
4 dependent mechanisms absent in stateless retrieval. Serial position effects in human memory
5 arise from differential rehearsal (Rundus, 1971) and working memory recency (Glanzer &
6 Cunitz, 1966; Atkinson & Shiffrin, 1968). RIF requires retrieval-dependent suppression of
7 competing representations (Anderson, 2003). Neither mechanism has a counterpart in standard
8 inverted indices or vector stores.
9
10
11
12
13
14
15
16

17 **Architectural Differences**

18

19 The divergence between BM25 and dense retrieval on encoding specificity connects to
20 the levels-of-processing framework (Craik & Lockhart, 1972). BM25 encodes at a surface level,
21 producing context-bound representations where retrieval depends on lexical match. Dense
22 retrieval encodes at a semantic level, producing representations that generalize across surface
23 variations. Both exhibit encoding specificity, but the grain of specificity differs.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 The distinction carries practical weight. Applications requiring robustness to query
43 rephrasing, such as conversational RAG or multilingual retrieval, may benefit from semantic
44 retrieval's reduced encoding specificity. Applications requiring precise contextual discrimination,
45 such as legal document retrieval where terminology carries specific meaning, may benefit from
46 the stronger encoding specificity of lexical approaches.
47
48

49 **Implications for Cognitively-Inspired Design**

50

51 For systems like HippoRAG (Gutierrez et al., 2024) and ARM (Bursa, 2026) that
52 explicitly pursue cognitive alignment, the results have specific implications. Competition-based
53 phenomena appear inherent to similarity-based retrieval and may not require special engineering.
54 If alignment on serial position or RIF is desired, fundamentally different mechanisms, such as
55 position-aware indexing or retrieval-dependent index modification, would likely be needed.
56
57
58
59
60

Limitations

Several limitations should be noted. First, the controlled synthetic corpus provides experimental precision but limited ecological validity; naturalistic corpus validation is needed. Second, we evaluate two architectures; hybrid, graph-based (Gutierrez et al., 2024; Edge et al., 2024), and adaptive systems may show different profiles. Third, corpus scale is modest, and scaling effects at realistic knowledge base sizes are unknown. Fourth, CAS normalization involves design choices warranting sensitivity analysis. Fifth, the mapping between cognitive tasks and retrieval tasks is approximate: documents are not episodic memories, and similarity-based ranking is not associative recall. Finally, human behavioral baselines from crowdsourced experiments would strengthen the comparison.

Future Work

Extensions include testing graph-based systems like HippoRAG for differential alignment; evaluating adaptive systems for emergent RIF; scaling corpus size; adding human baselines via Prolific or Amazon Mechanical Turk; extending modules to include the spacing effect (Cepeda et al., 2006), the testing effect (Roediger & Karpicke, 2006), levels-of-processing effects (Craik & Lockhart, 1972), and generation effects (Slamecka & Graf, 1978); and naturalistic corpus validation.

Conclusion

We introduced CogBench-RAG, a benchmark suite that evaluates retrieval-augmented generation systems against established human memory phenomena. Our evaluation indicates that BM25 and dense retrieval exhibit competition-based memory phenomena, including encoding specificity, the fan effect, and proactive/retroactive interference, while showing no evidence of serial position effects or retrieval-induced forgetting. The two architectures diverge on encoding

1
2
3 specificity, with lexical retrieval showing near-absolute context dependence and semantic
4 retrieval partially bridging contextual gaps. CogBench-RAG provides an extensible framework
5 for evaluating the cognitive properties of future retrieval architectures. Code and data are
6 available at <https://github.com/DestrierStudios/cogbench-rag>.
7
8

12 **Open Practices Statement** 13

14
15 All code, data, and materials for this study are publicly available at
16
17 <https://github.com/DestrierStudios/cogbench-rag> under an MIT license. The benchmark runs in a
18 Docker container for full reproducibility. No human participants were involved in this research.
19
20

21 **Data Availability Statement** 22

23
24 All data generated during this study are available in the public GitHub repository at
25
26 <https://github.com/DestrierStudios/cogbench-rag>. The repository includes all synthetic corpora,
27 query sets, benchmark code, and result files required to reproduce the reported findings.
28
29

30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 **References**
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6, 451–474.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
- Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2), 186–197.
- Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49(4), 415–445.
- Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(5), 1063–1087.
- Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of ICLR 2024.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press.
- Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (2nd ed.). Psychology Press.
- Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58(2), 97–105.
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.

- 1
2
3 Bursa, O. (2026). A dynamic retrieval-augmented generation system with selective memory and
4 remembrance. arXiv:2601.02428.
- 5
6 Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in
7 verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3),
8 354–380.
- 9
10 Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-
11 augmented generation. In Proceedings of AAAI 2024.
- 12
13 Chen, Y., Yan, L., Sun, W., Ma, X., Zhang, Y., Wang, S., Yin, D., Yang, Y., & Mao, J. (2025).
14 Improving retrieval-augmented generation through multi-agent reinforcement learning.
15 arXiv:2501.15228.
- 16
17 Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive
18 biases in moral decision-making. *Proceedings of the National Academy of Sciences*,
19 122(25), e2412015122.
- 20
21 Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory
22 research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684.
- 23
24 Ebbinghaus, H. (1885). *Über das Gedächtnis*. Duncker & Humblot.
- 25
26 Edge, D., Trinh, H., Cheng, N., et al. (2024). From local to global: A graph RAG approach to
27 query-focused summarization. arXiv:2404.16130.
- 28
29 Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory & Cognition*,
30 8(2), 157–173.
- 31
32 Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of
33 retrieval augmented generation. In Proceedings of EACL 2024.

- 1
2
3 Fountas, Z., et al. (2025). Human-inspired episodic memory for infinite context LLMs. In
4 Proceedings of ICLR 2025.
5
6 Germain, P. L., Sonrel, A., & Robinson, M. D. (2020). pipeComp: A general framework for the
7 evaluation of computational pipelines. *Genome Biology*, 21, 227.
8
9 Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal*
10
11 Learning and Verbal Behavior, 5(4), 351–360.
12
13 Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural
14 environments. *British Journal of Psychology*, 66, 325–331.
15
16 Gutierrez, B. J., et al. (2024). HippoRAG: Neurobiologically inspired long-term memory for
17 large language models. In NeurIPS 2024.
18
19 Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning
20 biases emerged in large language models but disappeared in ChatGPT. *Nature*
21
22 Computational Science, 3(10), 833–838.
23
24 Jiang, Z., et al. (2023). Active retrieval augmented generation. In Proceedings of EMNLP 2023.
25
26 Johnson, J., Douze, M., & Jegou, H. (2019). Billion-scale similarity search with GPUs. *IEEE*
27
28 *Transactions on Big Data*, 7(3), 535–547.
29
30 Kahana, M. J. (2012). Foundations of human memory. Oxford University Press.
31
32 Karpukhin, V., et al. (2020). Dense passage retrieval for open-domain question answering. In
33 Proceedings of EMNLP 2020.
34
35 Kim, S. H., et al. (2025). LLM reasoning does not protect against clinical cognitive biases.
36
37 medRxiv, 2025.06.22.25330078.
38
39 Kliegl, O., & Bauml, K.-H. T. (2021). Buildup and release from proactive interference.
40
41 Neuroscience & Biobehavioral Reviews, 120, 264–278.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Koo, R., et al. (2024). Benchmarking cognitive biases in large language models as evaluators. In
4
5 Findings of ACL 2024, 517–545.
6
7 Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In
8
9 NeurIPS 2020.
10
11 Liu, N. F., et al. (2024). Lost in the middle: How language models use long contexts.
12
13 Transactions of the ACL, 12, 157–173.
14
15 Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis.
16
17 Molecular Systems Biology, 15(6), e8746.
18
19 Ma, X., et al. (2021). A replication study of dense passage retriever. arXiv:2104.05740.
20
21 Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval.
22
23 Cambridge University Press.
24
25 Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories.
26
27 Journal of Experimental Psychology: General, 129(3), 361–368.
28
29 McGeoch, J. A. (1932). Forgetting and the law of disuse. Psychological Review, 39(4), 352–370.
30
31 McGaugh, J. L. (2000). Memory: A century of consolidation. Science, 287(5451), 248–251.
32
33 Muennighoff, N., et al. (2022). MTEB: Massive text embedding benchmark. arXiv:2210.07316.
34
35 Muller, G. E., & Pilzecker, A. (1900). Experimentelle Beitrage zur Lehre vom Gedachtnis.
36
37 Zeitschrift fur Psychologie, Supplement 1.
38
39 Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of
40 retrieval: A meta-analytic review. Psychological Bulletin, 140(5), 1383–1409.
41
42 Murdock, B. B. (1962). The serial position effect of free recall. Journal of Experimental
43 Psychology, 64, 482–488.

1
2
3 Osgood, C. E. (1949). The similarity paradox in human learning. *Psychological Review*, 56(3),
4
5 132–143.
6
7

8 Peysakhovich, A., & Lerer, A. (2023). Attention sorting combats recency bias in long context
9 language models. arXiv:2310.01427.
10
11

12 Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory &*
13
14 Cognition, 1, 19–40.
15
16

17 Radvansky, G. A., Spieler, D. H., & Zacks, R. T. (1993). Mental model organization. *Journal of*
18
19 Experimental Psychology: Learning, Memory, and Cognition, 19(1), 95–114.
20
21

22 Radvansky, G. A., & Zacks, J. M. (2014). Event cognition. Oxford University Press.
23
24

25 Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese
26 BERT-networks. In Proceedings of EMNLP 2019.
27
28

29 Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and
30 beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
31
32

33 Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, 17(3),
34
35 249–255.
36
37

38 Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental*
39
40 Psychology, 89(1), 63–77.
41
42

43 Saad-Falcon, J., et al. (2024). ARES: Automated evaluation framework for RAG systems. In
44 Proceedings of NAACL 2024.
45
46

47 Slamecka, N. J., & Graf, P. (1978). The generation effect. *Journal of Experimental Psychology:*
48
49 Human Learning and Memory, 4(6), 592–604.
50
51

52 Suri, G., et al. (2024). Do LLMs show decision heuristics similar to humans? *Journal of*
53
54 Experimental Psychology: General, 153(4), 1066–1075.
55
56
57
58
59
60

- 1
2
3 Teyler, T. J., & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral*
4
5 Neuroscience, 100(2), 147–154.
6
7 Thakur, N., et al. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of IR
8
9 models. In NeurIPS 2021 Datasets and Benchmarks Track.
10
11 Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.),
12
13 Organization of memory (pp. 381–403). Academic Press.
14
15 Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic
16
17 memory. *Psychological Review*, 80, 352–373.
18
19 Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, 64, 49–60.
20
21 Wang, W., et al. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression
22
23 of pre-trained transformers. In NeurIPS 2020.
24
25 Watkins, M. J., & Watkins, O. C. (1975). Buildup of proactive inhibition as a cue-overload
26
27 effect. *Journal of Experimental Psychology: Human Learning and Memory*, 1(4), 442–
28
29 452.
30
31 Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of*
32
33 *Psychology*, 55, 235–269.
34
35 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event
36
37 perception: A mind-brain perspective. *Psychological Bulletin*, 133(2), 273–293.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 February 18, 2026
4
5
6
7 Dear Editors,
8
9

10 I am writing to submit the manuscript entitled "Do Machines Remember Like We Do?
11 CogBench-RAG: A Cognitive Benchmark for Retrieval-Augmented Generation Systems" for
12 consideration as an Original Research Article in Behavior Research Methods.
13

14 Retrieval-Augmented Generation (RAG) systems are widely described using memory metaphors,
15 and recent architectures explicitly draw on cognitive neuroscience. Yet no evaluation framework
16 tests whether these systems actually exhibit well-characterized human memory phenomena. This
17 manuscript introduces CogBench-RAG, a benchmark suite that operationalizes five established
18 memory principles as controlled retrieval tasks: encoding specificity, the fan effect, proactive
19 and retroactive interference, serial position effects, and retrieval-induced forgetting. We evaluate
20 two architecturally distinct retrieval systems (BM25 and dense semantic retrieval) and quantify
21 their alignment with human memory patterns using a novel Cognitive Alignment Score.
22

23 Our results show that both systems exhibit three competition-based memory phenomena
24 (encoding specificity, the fan effect, and interference with the human-like retroactive-exceeds-
25 proactive asymmetry) while showing no evidence of serial position effects or retrieval-induced
26 forgetting. The two architectures diverge on encoding specificity: BM25 shows near-absolute
27 context dependence, while dense retrieval partially bridges contextual gaps. The null serial
28 position result establishes that standard retrieval systems do not contribute to the "lost in the
29 middle" effect documented in language models, localizing that bias to the generation stage.
30

31 We believe this work is well suited to Behavior Research Methods for several reasons. The paper
32 introduces a new methodological tool grounded in cognitive psychology that can be applied by
33 researchers studying human and machine cognition. The benchmark is open-source, fully
34 reproducible via Docker, and designed for extensibility. The work bridges cognitive science and
35 computer science in a way that should interest the broad readership of BRM, and it follows the
36 journal's emphasis on methods that readers can directly use in their own research.
37

38 This manuscript has not been published previously and is not under consideration elsewhere. All
39 code and data are publicly available at <https://github.com/DestrierStudios/cogbench-rag>. No
40 human participants were involved. The author declares no conflicts of interest and no external
41 funding.
42

43 Thank you for considering this submission. I look forward to hearing from you.
44

45
46
47
48 Sincerely,
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Nikhil Saxena
4
5 Northeastern University
6
7 Boston, MA 02115
8
9 saxena.ni@northeastern.edu
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only