I Found It in the Code, Science Proved It in the Lab: The Recency Bias That's Reshaping AI Search

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

Published: October 03, 2025 •

Canonical: https://metehan.ai/blog/i-found-it-in-the-code-science-proved-it-in-the-lab-the-recency-bias-thats-reshaping-ai-search/

In August 2025, I found something on ChatGPT's configuration files and identified a single line of code that explained many things:

```
use_freshness_scoring_profile: true
```

I wrote then: "ChatGPT actively prioritizes recent content over older material. Regular content updates aren't just good practice; they're essential for ChatGPT visibility." Here is the August 2025 post -> https://metehan.ai/blog/chatgpt-5-search-configuration/[1]

Today, I can tell you exactly how much this matters, because researchers just quantified it.

Figure

A team from Waseda University published a great study testing seven major AI models[2] (GPT-4o, GPT-4, GPT-3.5, LLaMA-3 8B/70B, and Qwen-2.5 7B/72B). They added artificial publication dates to search results and measured what happened.

The results validate everything I found in that configuration file and the numbers are interesting than I thought.

What I Found vs. What They Proved

My Discovery: The Configuration

Looking at ChatGPT's actual production settings, I found:

```
reranker_model: "ret-rr-skysight-v3"
use_freshness_scoring_profile: true
enable_query_intent: true
vocabulary_search_enabled: true
```

My conclusion: "That comprehensive guide you wrote in 2022? It might be losing ground to newer content, even if yours is more detailed."

Their Proof: The Numbers

The researchers took passages from TREC 2021 and 2022 test collections, added fake publication dates (nothing else changed same text, same quality), and watched AI models rerank them.

Every. Single. Model. Fell. For. It.

Here's what happened:

Metric	Best Case	Worst Case
Average year shift in top-10	+0.82 years (Qwen2.5-72B)	+4.78 years (LLaMA3-8B)
Largest single position jump	61 ranks (Qwen2.5-7B)	95 ranks (GPT-3.5-turbo)
Preference reversals	8.25% (Qwen2.5-72B)	25.23% (LLaMA3-8B)

Translation:

- Your top-10 results can shift by nearly 5 years just from timestamps
- Individual pieces of content can jump 95 positions
- 1 in 4 relevance decisions flip based solely on dates

The "Seesaw Effect": How Your Rankings Get Destroyed

The research revealed something fascinating they call the **"seesaw pattern"**and it perfectly explains what that freshness scoring profile actually does.

Imagine your search results as a seesaw with a pivot point in the middle:

Top 40 Positions: Systematically Younger

What happens: Content with recent dates (real or fake) consistently climbs here

By the numbers:

- Ranks 1-10: +0.8 to +4.8 years fresher (all models, both datasets)
- Ranks 11-20: +0.2 to +0.9 years fresher (statistically significant)
- Ranks 21-40: Still positive shifts, smaller magnitude

What this means: Even if you rank #1 based on content quality, a newer piece with worse content can overtake you.

⚖️ Ranks 41-60: The Pivot Point

What happens: Minimal movement, acts as the fulcrum

By the numbers:

- Some slight positive shifts in 41-50 band
- Some slight negative shifts in 51-60 band
- Mostly non-significant statistically

What this means: This is the "neutral zone" where freshness matters least.

Bottom 60: Systematically Older

What happens: Older-dated content sinks here, even when equally relevant

This freshness bias is also why citations in Google AI Mode move so much from one run to the next, which is why I tested the tools that track it in my guide to the best AI Mode rank trackers[3].

By the numbers:

- Ranks 61-70: -0.4 to -1.0 years older
- Ranks 71-80: -0.6 to -1.2 years older
- Ranks 81-90: -0.7 to -1.7 years older
- Ranks 91-100: -0.5 to -2.0 years older (most dramatic)

What this means: Older authoritative content gets systematically buried.

Real-World Impact: Three Scenarios

Scenario 1: Medical Content

What should happen: A landmark 2018 study with 10,000 participants and peer review should rank highly.

What actually happens: A preliminary 2024 blog post with 50-person sample and no peer review ranks higher just because it's newer.

The numbers: The 2018 study could drop 40-60 positions purely from its date.

Scenario 2: Technical Documentation

What should happen: The definitive 2020 guide with 5,000 verifications and community vetting should be authoritative.

What actually happens: A 2024 unverified blog post ranks higher.

The numbers: Up to 25% chance the AI "prefers" the newer, worse content.

Scenario 3: Academic Research

What should happen: Foundational papers from 2015-2020 should remain authoritative reference material.

What actually happens: Recent commentary pieces with no original research rank higher.

The numbers: Top-10 can shift 1-5 years newer, systematically demoting classics.

The Configuration + Research = Complete Picture

Let me show you how my configuration discovery and their research fit together:

1. The Reranker (`ret-rr-skysight-v3`)

What I found: ChatGPT uses a sophisticated reranking model that processes search results post-retrieval.

What research adds: This isn't unique to ChatGPT **all listwise rerankers** exhibit this bias. It's architectural, not implementation-specific.

New insight: The Skysight-v3 model likely has temporal bias **built into its training**, not just as a configuration parameter.

2. Freshness Scoring

What I found: `use_freshness_scoring_profile: true` is always on.

What research adds: The effect magnitude is 1 to 5 years of shift in top results.

New insight: This isn't a minor ranking signal. It's **dominant enough to override content quality signals**.

3. Query Intent Detection

What I found: `enable_query_intent: true` means ChatGPT analyzes what you're actually trying to accomplish.

What research adds: Intent detection **doesn't adjust for temporal appropriateness**. Historical queries get the same freshness bias as news queries.

New insight: A query like "causes of World War I" shouldn't prioritize 2024 content, but it does. The intent detection isn't temporally aware.

4. Vocabulary Search

What I found: `vocabulary_search_enabled: true` with fine-grained filtering rewards technical terminology.

What research adds: Even content with **perfect vocabulary** loses to newer content with **worse vocabulary** up to 25% of the time.

New insight: Technical accuracy Original configuration analysis: Inside ChatGPT's GPT 5 Search Configuration[1]

- Academic research: "Do Large Language Models Favor Recent Content? A Study on Recency Bias in LLM-Based Reranking" by Fang et al., Waseda University, 2025[2]

Recency bias interacts directly with model training windows — see the fact-checked LLM knowledge cutoff dates[4] and how AI platforms choose citations[5].

Also related: what prompt engineering means in the context of AI[6].

Recency bias is one of the strongest AEO signals we can prove in code and in the lab. AI citations in ChatGPT, Perplexity and Gemini lean toward recent pages in ways answer engine optimization programs have to design around.

REFERENCES & SOURCES

[1] **https://metehan.ai/blog/chatgpt-5-search-configuration/** — https://metehan.ai/blog/chatgpt-5-search-configuration/

[2] **published a great study testing seven major AI models** — https://arxiv.org/abs/2509.11353

[3] **the best AI Mode rank trackers** — https://metehan.ai/articles/best-ai-mode-rank-tracker

[4] **LLM knowledge cutoff dates** — https://metehan.ai/articles/llm-knowledge-cutoff-dates/

[5] **how AI platforms choose citations** — https://metehan.ai/articles/ai-platforms-citations-patterns/

[6] **what prompt engineering means in the context of AI** — https://metehan.ai/articles/what-is-prompt-engineering-in-the-context-of-ai/