Individual Text Corpora Predict User-Specific Knowledge: Benchmarks of Individualized Knowledge Simulation

Christoph Wigbels[1,†] **Ali Abusaleh**[2,†] **Markus T. Jansen**[1]

Alexander Mehler[2] **Manuel Schaaf**[2] **Markus J. Hofmann**[1]

[†] Equal contribution [1]Bergische Universität Wuppertal [2]Goethe-Universität Frankfurt

{wigbels, mhofmann, mjansen}@uni-wuppertal.de

{a.abusaleh, mehler, manuel.schaaf}@em.uni-frankfurt.de

Abstract

This study examines whether individual text corpora (ICs) from search histories can be used to simulate individual knowledge. We collected ICs from 316 adults, who answered 36 multiple-choice knowledge items, and compared several large language models (LLMs) on this task, of which only Qwen3-1.7B proved viable. After task-specific fine-tuning via Low-Rank Adaptation (LoRA), Qwen3-1.7B outperformed both participants and a representative German norm sample on publicly available items. On non-public questions, however, the LLM performed worse than our participants, suggesting possible training data contamination for the public questions. When integrating ICs into retrieval-augmented generation to predict individual responses, LLM-participant Match accuracies significantly exceeded chance, which demonstrates a detectable individual knowledge signal. The probabilities assigned to the participants' answers were, however, low and far below the probability of correct answers, indicating poor calibration toward individual response patterns. Knowledge-gap prediction was suboptimal, though it improved for corpora exceeding five million tokens. We discuss our entropy-based evaluation benchmarks as calibration indices for individualized knowledge simulation.

1 Introduction

Large language models (LLMs) store substantial amounts of factual knowledge within their parameters. However, this knowledge is static and reflects aggregate patterns in training corpora rather than the knowledge of specific individuals (Petroni et al., 2019). For applications such as adaptive tutoring or personalized diagnostics, it is not sufficient to know what is generally true; instead, a system must model what a particular user knows or does not know to allocate instructional effort effectively (Corbett and Anderson, 1995; VanLehn, 2011). The present study addresses this challenge by grounding LLM outputs in individual text corpora (ICs) derived from internet search histories. Psychologically, individual differences in intellectual abilities are commonly distinguished into fluid intelligence (Gf) and crystallized intelligence (Gc). While Gc is viewed as the ability to use factual and conceptual knowledge (crystallized knowledge; CK), Gf denotes the capacity to solve novel problems independent of Gc (Cattell, 1963). Over time, Gf is invested in domain-specific learning, yielding highly individualized knowledge profiles shaped by interests and experience, a component described as the "dark matter" of adult intelligence (Ackerman, 1996, 2000). Simulating what a particular user knows therefore requires behavioral traces of this individual investment process. Internet search behavior provides a promising signal: each visited page reflects active engagement with specific content and contributes to a cumulative personal reading history (Marchionini, 2006; Dumais et al., 2003; Teevan et al., 2007). We refer to the factual information encoded in LLM weights as parametric knowledge and use retrieval-augmented generation (RAG) as our computational framework. Standard RAG grounds model outputs in documents retrieved from a shared knowledge base, separating parametric from contextual information (Lewis et al., 2020; Gao et al., 2024). To evaluate this approach, we collected Google search histories from 316 participants (mean IC size: 3.2 million tokens) and constructed individual corpora via web scraping of visited URLs. Each participant completed 36 multiple-choice knowledge items. To address potential training-data contamination in LLM evaluation (Deng et al., 2024; Zhao et al., 2025), we distinguished between 12 publicly available BEFKI GC-K items (Schipolowski et al., 2013), potentially included in pretraining data, and 24 newly developed items. An initial comparison showed that only Qwen3-1.7B reached satisfactory per-

formance; GerPT2 and GPT-2 failed to exceed chance. We therefore evaluate Qwen3-1.7B in three stages: (1) zero-shot parametric knowledge, (2) task-specific fine-tuning via LoRA on 129 training items (Zimny et al., 2024), and (3) IC-based RAG. Our evaluation addresses two complementary questions that we keep terminologically distinct throughout the paper. First, we assess factual correctness, that is, whether the selected option matches the keyed correct answer, which we report for the model and for the participants alike. Second, we assess how well the model simulates a specific participant, that is, whether it accurately reproduces the answer this person actually chose, irrespective of whether that answer is correct, and whether it predicts which items this person answered correctly or incorrectly. We reserve the term accuracy for this second, person-oriented perspective and evaluate it using a discrete Match accuracy, a probabilistic log-loss metric, and crystallized-knowledge accuracy (CK-accuracy).

2 Related Work

2.1 Individual Text Corpora and Episodic Retrieval

While Pennebaker and King (1999) provided early correlational research examining personal diary entries, later work captured interindividual variability in personality using a common language model (Eichstaedt et al., 2021). More recent research introduced the IC approach, based on the idea that a text corpus can serve as a sample of a person's experience. A single language model is trained for each person, so that training parallels memory consolidation and inference parallels retrieval (Hofmann et al., 2024). Hofmann et al. (2020) first demonstrated this approach by constructing ICs from individuals' everyday reading behavior and training language models directly on these corpora, showing that such individually trained models predict word fixation times in an eye-tracking paradigm more accurately than models trained on standard corpora. Hofmann et al. (2024) then scaled this approach to web-scale behavioral data, constructing ICs from Google search histories and showing that IC-derived similarity measures predict psychological traits such as openness to experience and intellectual engagement in held-out samples. Prior studies trained language models directly on ICs, implicitly modeling semantic long-term memory. We instead use RAG to dynamically retrieve relevant IC fragments per query, shifting from consolidated semantic memory to episodic retrieval (Dong et al., 2025). This allows us to test whether person-specific corpora support situation-dependent knowledge retrieval.

2.2 RAG, Knowledge Probing, and Personalized LLMs

Our work connects to three related research areas in natural language processing (NLP). First, knowledge probing studies have shown that pretrained language models encode substantial factual knowledge that can be accessed through cloze-style queries, while also revealing systematic gaps and inconsistencies in this knowledge (Petroni et al., 2019). Second, RAG mitigates these limitations through external retrieval (Lewis et al., 2020). Third, recent research on personalized LLMs investigates how user-specific data, such as interaction logs, preferences, or personal documents, can be used to adapt model behavior to individual users (Zhang et al., 2025). IC-based RAG sits at the intersection of these three lines of work. Like knowledge probing, it is concerned with what a model knows; like RAG, it relies on external retrieval to support generation; and like personalization research, it incorporates user-specific data. Crucially, however, it differs in its evaluation target: instead of optimizing for general factual correctness or user preference, we evaluate whether the model can reproduce individual knowledge. This shifts the focus from generic knowledge augmentation to the simulation of person-specific cognitive states.

3 Methods

3.1 Participants and Knowledge Assessment

The final sample comprised 316 adults (71.4% female; 65.6% aged 18–25, 27.3% aged 26–35) with diverse educational backgrounds (63.0% Abitur, 16.9% university degree). Participants were recruited via university flyers, the SONA system, the PsyWeb panel[1], and personal recruitment. Eligibility required fluency in German, active use of a Google account for at least one year, and participation in a 60–90-minute online survey covering knowledge, personality, interests, fluid intelligence, and demographic information. Participants received either course credits or 18 €. Participants were excluded if they did not provide a

[1] https://psyweb.uni-muenster.de

valid search history file, if their IC contained fewer than 2,500 stemmed word types, or if they failed all control questions. Crystallized knowledge was assessed using 36 multiple-choice items. These included the 12 items from the BEFKI GC-K short scale (Schipolowski et al., 2013) and 24 new items. In addition, 129 items (Zimny et al., 2024)[2] were used for task-specific fine-tuning and are distinct from the 36 evaluation items. Each item consisted of a question stem and four answer options, with exactly one correct answer; an example item is shown in Appendix A.

3.2 Individual Text Corpora and Web Scraping

ICs were constructed from participants' search histories. Participants exported their data via Google Takeout; extracted URLs were then scraped using a two-pass pipeline (HTTP extraction followed by JavaScript rendering) with rotating proxies and PDF support. A cleaning filter removed boilerplate content and non-German text (sentences with fewer than three German stop words were excluded).

We next evaluated how pretrained language models perform on the task before adapting them to the specific question format and integrating participant-specific corpora via RAG.

3.3 Task-Specific Fine-Tuning of Pretrained Language Models

We investigated the capability of LLMs to process and answer multiple-choice questions within a structured evaluation framework. Our methodology comprised three primary phases: (1) establishing baseline performance through zero-shot evaluation, (2) adapting the models to the specific question format using Low-Rank Adaptation (LoRA) (Hu et al., 2021) rather than full-parameter fine-tuning, and (3) simulating participant-specific knowledge by integrating personal corpora via RAG during the question-answering process.

3.4 Baseline and Task-Specific Fine-Tuning

Baseline Evaluation We first evaluated several language models on the multiple-choice task in a zero-shot setting to establish a baseline. The evaluated models included GPT-2 (Radford et al., 2019), GerPT2 (Minixhofer, 2020), and Qwen3-1.7B (Qwen Team, 2025). Performance correctness was calculated based on an exact match between the model's predicted option[3] and the ground-truth answer.

```
Please answer with one of the
    options in the bracket.
Write the answer
in between <answer></answer>.

### Question:
{question text with options A–D}

### Response:
<answer>
{correct letter}. {correct text}
</answer>
```

Listing 1: Prompt for QA Fine-tuning

Task-Specific Fine-Tuning To adapt the model to the structured multiple-choice format while keeping computational costs low, we applied parameter-efficient fine-tuning using LoRA. Rather than updating all model parameters, LoRA introduces trainable low-rank adapter matrices into selected layers while keeping the original weights frozen.

The fine-tuning dataset comprised 129 multiple-choice items formatted identically to the evaluation items. Each prompt included a question stem and four answer options, with the correct answer encoded as a letter (A–D). The prompt format enforced a standardized output structure compatible with downstream evaluation (see Listing 1).

Fine-tuning was performed with a LoRA rank of $r = 8$, a scaling factor $\alpha = 16$, and a dropout rate of 0.05 for regularization. Adapters were injected into multiple projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`) to maximize adaptation capacity. Training used the 32-bit Paged AdamW optimizer (`paged_adamw_32bit`) with a learning rate of 2×10^{-4}. Models were trained for 10 epochs with a per-device batch size of 1 and 2 gradient accumulation steps, along with 10 warmup steps to stabilize training.

After adapting the model to the task format, we incorporated participant-specific information via RAG.

[2] https://osf.io/u68nk/files

[3] Given the small size of the test dataset, manual post-processing and extraction of the answers were sufficient for evaluation.

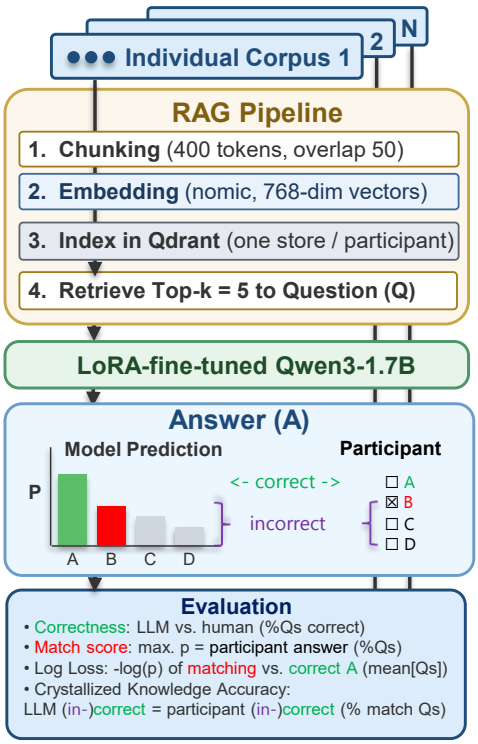

Figure 1: Study Overview: Language Modeling Pipeline and Evaluation

3.5 RAG Pipeline

Each corpus was segmented into chunks of 400 tokens with an overlap of 50 tokens to preserve contextual continuity. These chunks were embedded using the `nomic-embed-text` model (Nussbaum et al., 2025), producing 768-dimensional vector representations, and stored in a Qdrant vector database (Qdrant Team, n.d.). At inference time, the system retrieves the top $k = 5$ most semantically relevant chunks for a given question. These retrieved text fragments are then incorporated into the model prompt, grounding the generated response in the participant's personal corpus.

These modeling steps yield both discrete predictions and full probability distributions over answer options. We evaluate model performance using complementary metrics that capture both correctness and probabilistic calibration (see Figure 1).

Inference Prompt Design While fine-tuning, the model was conditioned on the multiple-choice format (Listing 1). At evaluation time, the prompt is augmented with retrieved participant context and an explicit instruction to rely on it (Listing 2). The instruction directs the model to answer *based only* on the provided search-history context, rather than on its parametric knowledge, with a fallback clause

```
You are a specialized assistant.
You MUST answer the question
based ONLY on the provided search
history context. If the answer is
not in the context, use your best
judgment based on the participant
interests shown in the history.

### SEARCH HISTORY CONTEXT:
{retrieved participant search
    history}
## QUESTION:
{question text}
A. {option A}
B. {option B}
C. {option C}
D. {option D}
### Response:
<answer>
```

Listing 2: Inference prompt combining Context and Question

that permits reliance on the participant's demonstrated interests when the answer is not directly retrievable.

3.6 Evaluation

We evaluated model performance along four complementary dimensions.

Answer correctness. We first measured standard multiple-choice factual knowledge correctness by comparing the model's predicted option with the keyed correct answer.

Match accuracy. To assess how well the RAG model reproduces individual responses, we computed the Match accuracy, defined as the proportion of items for which the model's most likely option matches the participant's choice.

Log loss. To evaluate the full predicted response distribution, we computed multiclass log loss. For each item i, the model produces a probability vector $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,4})$, and the participant's response is represented as a one-hot vector $\mathbf{y}_i$. This allows us to interpret the model's uncertainty in information-theoretic terms: log loss measures how much probability mass the model assigns to the observed response and is equivalent to the cross-

entropy

$$H(\mathbf{y}_i, \mathbf{p}_i) = -\sum_{j=1}^{4} y_{i,j} \log p_{i,j}, \qquad (1)$$

which, for one-hot targets, is equivalent to the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(\mathbf{y}_i \,\|\, \mathbf{p}_i) = \sum_{j=1}^{4} y_{i,j} \log \frac{y_{i,j}}{p_{i,j}}. \qquad (2)$$

Averaging across N items yields the log-loss score,

$$\mathrm{LogLoss} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{i,j_i^*}, \qquad (3)$$

where j_i^* denotes the option chosen by the participant. Log loss is measured in natural units of information (nats). Lower values indicate better alignment between predicted probabilities and observed responses. For instance, the chance-level baseline corresponds to a uniform distribution ($-\log 0.25 = 1.39$ nats).

CK-accuracy. While the Match accuracy asks whether the maximum probability answer of the LLM is the participant's selected answer, we were also interested in a measure that asks whether the LLM can accurately predict whether a participant answered an item correctly or incorrectly. Therefore, we evaluated knowledge-gap prediction using crystallized-knowledge accuracy (CK-accuracy), defined as the proportion of items for which the model accurately predicts whether a participant answered correctly or incorrectly.

Model	Zero-shot	Fine-Tuned
GerPT2	0%	4%
GPT-2	0%	22.7%
Qwen3-1.7B	81.1%	100%

Table 1: LLM Selection Based on Correctness

All metrics were averaged across participants and compared to their respective baselines using one-sample *t*-tests.

4 Results

4.1 LLM Selection

Models were evaluated on a held-out test set of 22 multiple-choice items (Zimny et al., 2024) that were not included in the fine-tuning corpus and are distinct from the 36 evaluation items used in the main analysis (Table 1). The results indicate that, while smaller models (GPT-2, GerPT2) showed modest improvements (Minixhofer, 2020; Radford et al., 2019), the Qwen3-1.7B model reached 100% correctness on the held-out set after fine-tuning (Qwen Team, 2025). Correctness before fine-tuning ranged from 0% to 81.1%; the gains after fine-tuning demonstrate the impact of task-specific adaptation. However, in this initial selection set, the correct option was always A, so a model could in principle have reached a perfect score simply by always producing the letter A, reflecting label bias rather than genuine knowledge. We therefore (i) treated this comparison only as a preliminary viability check for model choice and (ii) retrained Qwen3-1.7B for the final evaluation on items with randomized correct-answer positions, so that no constant-response strategy could inflate its scores.

4.2 LLM Versus Human Answer Correctness

Participant correctness (mean proportion correct per participant) was compared to the fixed item-level correctness of the fine-tuned Qwen3-1.7B model (Figure 2). One-sample *t*-tests were used to test whether the participant mean differed from the constant LLM correctness (Table 2). Across all 36 items, the LLM achieved higher correctness than participants. However, this effect was confined to the 12 BEFKI GC-K core items, where the model substantially outperformed the participant sample (81% vs. 63%). On the 24 extended items, participants outperformed the LLM (65% vs. 60%). To contextualize these results, we compared both groups to a nationally representative German norm sample (N = 1,134; Schipolowski et al., 2013). On the core items, the norm sample scored .59 (SD = .22), which is below our participant sample (.63) and the LLM (.81).

4.3 Individualized Knowledge Simulation

IC-based RAG Match accuracies and log loss were tested against chance (0.25 for Match accuracy; 1.39 nats for log loss) using one-sample *t*-tests (see Table 2). Match accuracies significantly exceeded chance across all item sets, reaching an overall mean of .31 against the .25 chance level ($\Delta = +.06$, $d = 1.06$), with the strongest effect for core items (.36) and a moderate effect for extended items (.29). However, this improvement in dis-

Metric	Item set	M (SD)	Baseline	t	p	d	95% CI
Participant correctness	All (36)	.64 (.14)	.67[a]	-3.79	$< .001$	-0.21	$[-.044, -.014]$
	Core (12)	.63 (.16)	.81[a]	-20.12	$< .001$	-1.13	$[-.204, -.168]$
	Extended (24)	.65 (.14)	.60[a]	6.10	$< .001$	0.34	$[.034, .066]$
Norm correctness[e]	Core (12)	.59 (.22)	.81[a]	-33.67	$< .001$	-1.00	$[-.233, -.207]$
Match Accuracy	All (36)	.31 (.06)	.25[b]	18.84	$< .001$	1.06	$[.057, .070]$
	Core (12)	.36 (.10)	.25[b]	19.50	$< .001$	1.10	$[.102, .125]$
	Extended (24)	.29 (.07)	.25[b]	9.52	$< .001$	0.54	$[.031, .047]$
Log loss: match (nats)	All (36)	6.29 (0.72)	1.39[c]	121.06	$< .001$	6.81	$[4.820, 4.978]$
	Core (12)	6.26 (1.19)	1.39[c]	72.69	$< .001$	4.09	$[4.742, 5.005]$
	Extended (24)	6.30 (0.82)	1.39[c]	105.98	$< .001$	5.96	$[4.821, 5.003]$
Log loss: correct (nats)	All (36)	1.56 (0.26)	1.39[c]	12.22	$< .001$	0.69	$[.148, .205]$
	Core (12)	0.99 (0.29)	1.39[c]	-23.64	$< .001$	-1.33	$[-.425, -.359]$
	Extended (24)	1.86 (0.35)	1.39[c]	23.96	$< .001$	1.35	$[.435, .512]$
CK-accuracy	All (36)	.54 (.08)	.64[d]	-22.67	$< .001$	-1.28	$[-.107, -.090]$
	Core (12)	.54 (.14)	.63[d]	-11.41	$< .001$	-0.64	$[-.103, -.073]$
	Extended (24)	.55 (.09)	.65[d]	-21.48	$< .001$	-1.21	$[-.112, -.094]$

Table 2: One-Sample t-tests Comparing Evaluation Metrics Against Their Respective Baselines

Note. $N = 316$ for all metrics except norm correctness. Core = BEFKI GC-K short form (Schipolowski et al., 2013); Extended = newly developed items not publicly available. Values in parentheses are standard deviations. 95% CIs refer to the mean difference $M -$ Baseline. All t-tests use $df = 315$ except norm correctness, which uses $df = 1{,}133$. Log loss: correct = log loss computed against the keyed correct answer rather than the participant's chosen option. [a]LLM correctness (item-level proportion correct). [b]Four-option chance level. [c]Uniform distribution: $-\ln(0.25) = 1.39$ nats. [d]Majority-class baseline (overall participant correctness per item set). [e]Norm correctness = correctness for a nationally representative German norm sample ($N = 1{,}134$; Schipolowski et al., 2013).

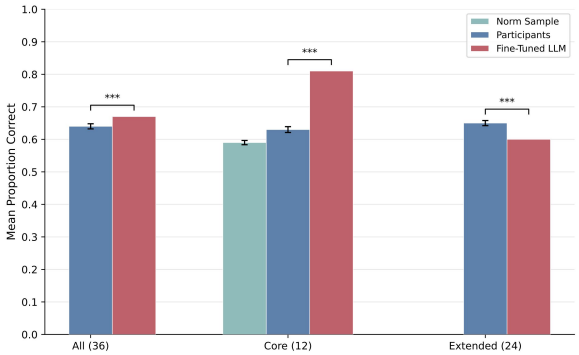

Figure 2: Mean Proportion Correct for the Representative Norm Sample, Study Participants, and the Fine-Tuned LLM by Item Set

Note. Error bars represent ± 1 *SE*. $N = 316$; norm data from a nationally representative German sample ($N = 1{,}134$; Schipolowski et al., 2013), available for core items only. LLM correctness is a fixed item-level proportion (no error bar). $^{***}p < .001$.

crete matching did not translate into well-calibrated probabilistic predictions. Log loss was consistently higher than chance across all item sets, indicating that the model assigned low average probabilities to the responses actually chosen by participants. This discrepancy reflects a systematic pattern: while the model often selected the same answer as the participant, it did so with highly concentrated probability distributions. The low mean normalized entropy (.10) further indicates that the model failed to represent uncertainty across alternative options.

A complementary analysis computing log loss against the keyed correct answer revealed a dissociation: for core BEFKI items, log loss fell below baseline, whereas for extended items it was clearly above baseline.

Finally, when evaluating whether the model can predict which items a participant answered correctly or incorrectly, CK-accuracy fell below the majority-class baseline across all item sets (see Table 2), indicating that the model did not reliably distinguish between items that participants knew and those they did not.

4.4 Corpus Size as a Moderator

We further explored the association between corpus size ($\log_{10}$ token count) and evaluation metrics with partial correlations, thus controlling for participant correctness. Corpus size was not significantly associated with any evaluation metric (all ps $> .13$). However, a different pattern emerged among participants with larger corpora. In the top-30% subsample ($n = 95$; ≥ 5.02 million tokens), corpus size was positively associated with CK-accuracy, $r(93) = .36$, $p < .001$ (see Figure 3) and with participant correctness, $r(93) = .25$, $p = .013$. The association between corpus size and CK-accuracy

remained significant after controlling for participant correctness, partial $r(92) = .28$, $p = .006$. These findings indicate that prediction quality improves once a sufficient corpus size threshold is reached.

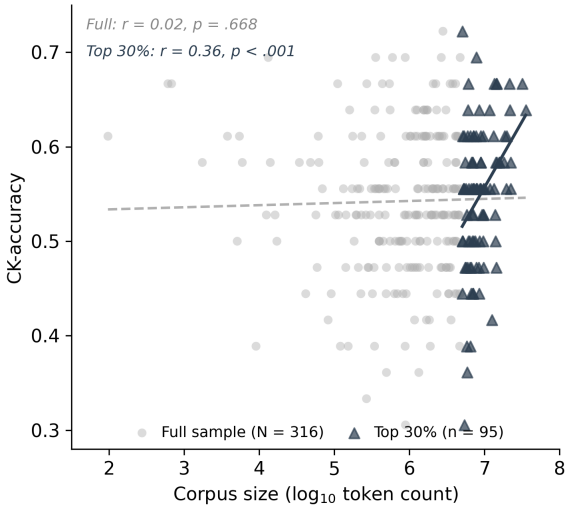

Figure 3: Corpus Size and CK-accuracy for the Full Sample and the Top-30% Subsample

Note. Pearson correlation between corpus size ($\log_{10}$ token count) and CK-accuracy. Gray circles (dashed regression line) = full sample ($N = 316$); dark triangles (solid regression line) = top-30% subsample ($n = 95$; ≥ 5.02 million tokens). $^{*}p < .05$. $^{***}p < .001$.

5 Discussion

This study examined whether ICs derived from Google search histories can serve as a proxy for person-specific crystallized knowledge. As reported above, only Qwen3-1.7B proved viable after task-specific fine-tuning; all subsequent analyses relied exclusively on this model.

A notable dissociation emerged when comparing model performance across different item types. On the publicly available BEFKI core items, the LLM outperformed human participants (81% vs. 63% correctness), whereas on the newly developed extended items, it slightly underperformed (60% vs. 65% correctness). This pattern is consistent with the problem of training-data contamination in standardized assessments, and it simultaneously demonstrates that human crystallized intelligence still surpasses parametric world knowledge when confronted with genuinely novel problems. In this regard, the results underscore the urgent need for contamination-controlled benchmarks in LLM evaluation, as has been argued in recent work (e.g., Jacovi et al., 2023). Performance on public items likely reflects parametric retrieval rather than genuine inference, which suggests that standard benchmark scores may substantially overestimate a model's actual reasoning capacity. The 18-percentage-point advantage on the public BEFKI core items, compared with the model's underperformance on the contamination-controlled extended items, illustrates this concern concretely and is consistent with the view that novel, previously unseen evaluation sets are necessary when claims about model reasoning are to be assessed.

Regarding the central question of individualized knowledge simulation, IC-based RAG yielded Match accuracies, that is, agreement with the answers the participants actually chose, consistently above the chance level of .25 (overall .31, $\Delta = +.06$, $d = 1.06$). To our knowledge, this result constitutes the first empirical demonstration that person-specific text corpora derived from search histories carry a detectable signal of individual knowledge. The large effect size primarily reflects low between-person variance rather than a substantial practical gain. Moreover, the log-loss analysis reveals a fundamental limitation that correctness metrics alone would obscure: the observed log-loss values indicate that the model fails to reflect the distribution of incorrect options a participant might actually consider, though it predicts the correct option with extreme confidence.

This dissociation points to a critical optimization gap: future individualized models must shift from answer-oriented to person-oriented training, for instance by incorporating participants' actual response distributions during fine-tuning. Furthermore, CK-accuracy fell below the majority-class baseline (.54 vs. .64), indicating that the model does not yet capture individual variations in knowledge gaps. This asymmetry, whereby the model performs above chance for what participants know but fails for what they do not know, offers at least two interpretations. First, the structure of the data itself plays a role: search histories record moments of active information seeking rather than failed recall, such that gaps manifest as silence within the IC. Second, a theoretical explanation rooted in PPIK (Process, Personality, Interests, and Knowledge) theory (Ackerman, 1996, 2000) suggests that digital behavior primarily reflects areas of sustained intellectual engagement; the absence of search activity is therefore only a weak indicator of missing knowledge, as a person may possess knowledge acquired through non-digital channels.

Crucially, the predictive signal scaled with data volume: only in the top-30% subsample (≥ 5.02 million tokens) did corpus size significantly predict CK-accuracy, even after controlling for participant knowledge. This suggests that the approach is currently data-limited rather than fundamentally flawed, and the threshold at approximately five million tokens establishes a concrete design parameter for future personalized RAG systems.

6 Conclusions and Outlook

This study demonstrates that ICs from search histories carry a measurable signal of person-specific crystallized knowledge when combined with RAG. Match accuracies establish a proof of concept, but log-loss analysis identifies calibration as the primary bottleneck. The corpus-size effect locates the current bottleneck in data volume rather than in the approach itself. Together, our benchmarks allow for a clear developmental trajectory: shifting from answer-oriented to person-oriented optimization, enriching ICs with behavioral metadata such as dwell time and re-visitation frequency, and ensuring sufficient corpus size above the approximately five-million-token threshold identified here. In particular, minimizing log loss against participants' actual choices could directly incentivize calibrated probability estimates. From a cognitive perspective, we interpret RAG-based retrieval as an episodic memory retrieval process (Dong et al., 2025), though individuals likely also differ in semantic long-term memory. Our benchmarks may also help investigate the inability of LLMs to forget pretraining knowledge (Jang et al., 2023; Yao et al., 2024; Blanco-Justicia et al., 2025). Given the corpus-size effect, future work should include English-language websites.

A first and immediate extension addresses the scope of the present evaluation, which rested on a single model. Model size is only one of several levers for individualizing language models (Zhang et al., 2025). Because the poor calibration we observed coincided with the strong parametric factual knowledge of Qwen3-1.7B, smaller models such as Qwen3-0.6B, with less memorized world knowledge to fall back on, are a natural next comparison, consistent with evidence that smaller language models can align more closely with human knowledge distributions than larger ones (He-Yueya et al., 2024). Such small models may leave more room for the individual signal to shape the answer distribution, which would speak to whether the calibration gap is tied to model scale. In parallel, the person-oriented objective sketched above can be realized through per-participant fine-tuning that writes each individual corpus directly into the model weights with a weight-decomposed low-rank adapter (Liu et al., 2024), optimizing the model to reproduce a specific person rather than the keyed correct answer. We are currently pursuing both directions, and report an exploratory comparison of further models under the identical pipeline in Appendix B. A further direction concerns the integration of interests as a complementary individualizing signal. Because knowledge gaps remain largely invisible to IC-based retrieval, as reflected in the below-baseline CK-accuracy, the corpus alone says little about what a person does not know. Interests, which the present survey already assessed with the FIFI-K, a short inventory of leisure interests (Nikstat et al., 2018), offer a principled if broad proxy: within the PPIK framework, sustained interest drives the acquisition of domain knowledge, so that low interest in a domain is itself informative about likely gaps. As leisure interests capture only one facet of the interests that drive knowledge acquisition, a domain-specific or epistemic interest measure would sharpen this signal further. Combining the retrieved corpus content with an explicit interest profile could thus improve knowledge-gap prediction precisely where the corpus fails to provide a signal. Ultimately, successful individualized simulation could enable tutoring agents to select instructional texts that minimize the log-loss gap between model predictions and correct answers.

Limitations

Notwithstanding the previously mentioned findings, several limitations must be critically noted. First, the sample was predominantly young, female, and highly educated (65.6% aged 18–25; 71.4% female; 63.0% Abitur), which may limit the generalizability of the results to demographic groups whose search behavior and knowledge profiles differ systematically. Second, the study is language-specific at this stage: all knowledge items were administered in German, all ICs were filtered to retain German-language content, and the stopword-based filter may have excluded relevant multilingual material. Also, it remains unclear whether the approach generalizes to other languages or to multilingual users whose search histories span mul-

tiple languages. Third, the construction of the ICs relied exclusively on Google search histories. Although this source provides a scalable form of data donation, it captures only one outlet of intellectual engagement and excludes other potentially informative digital traces, such as e-book annotations or YouTube watch history. The latter is also accessible via Google Takeout and could be incorporated in future work through auto-generated captions. Fourth, the mapping from the IC to what a person actually knows could be indirect. A visited page may index interest and information seeking rather than secured knowledge, so that the presence of a topic in the corpus does not imply that the person mastered it, also reading a text does not guarantee that its content was understood or retained. Beyond this, web search itself can function as a form of external, or transactive, memory: as people come to expect continuous online access to information, they tend to remember where to find it rather than the content itself (Sparrow et al., 2011). A search may therefore not reflect internalized knowledge at all, and it can instead represent precisely the content a person has offloaded to the web, which further weakens the corpus-knowledge link. The IC is therefore a proxy for exposure and engagement rather than for knowledge itself, complementing the reverse point discussed above. In line with this, the overlap between the corpora and the tested items is uneven and follows the searchability of a topic rather than its curricular importance: everyday knowledge that people routinely look up, for instance the Zugspitze, the legal form GmbH, the founding of the German Reich, or Nietzsche, tends to be well represented, whereas canonical school knowledge such as mitosis or the mitochondrion appears far less often. Notably, coverage was not higher for the publicly available BEFKI core items than for the non-public extended items, so that coverage does not simply track an item's difficulty or public availability. Fifth, only a single model architecture (Qwen3-1.7B) and one retrieval configuration were evaluated in the main RAG setting. The model's high answer correctness was accompanied by high overconfidence, which likely contributed to the poor log-loss performance. This highlights a central limitation of the present approach: strong item-level correctness does not necessarily imply well-calibrated simulation of individual response distributions. Because no alternative model sizes, calibration procedures, or retrieval strategies were compared, it remains unclear whether the observed trade-off is specific to the present implementation or reflects a more general limitation of IC-based answer simulation.

Ethical Considerations and Data Availability

The data collection for the overarching research project was approved by the ethics committee of Bergische Universität Wuppertal. All participants provided informed consent, including explicit agreement to share their Google search history data for scientific purposes. We invested considerable effort into anonymizing the individual text corpora. The ICs do not contain person-identifying information, raw URLs, or precise timestamps; only coarse timestamps relative to assessment time are retained. Because ICs represent a subset of the information collected during commercial web tracking, participant identifiability should be lower than with standard tracking data (Deußer et al., 2020). As web search data are already used extensively for commercial purposes, we consider it an ethical necessity to lead an open scientific discussion about the possibilities and limitations of such data for psychological research. In contrast to commercial objectives, our work aims to improve individualized psychodiagnostics and adaptive tutoring. Nevertheless, the risk of re-identification from large personal text corpora cannot be fully excluded. For this reason, the ICs are not publicly released at this time. Instead, we plan to provide them through the GESIS data archive under controlled access, so that secondary analyses remain restricted to scientific purposes covered by the participants' informed consent. Because the residual risk of re-identification should be measured rather than merely assumed, we further plan a shared task on person identifiability. This shared task rests on the same safeguards as the present study, that is, explicit consent to data donation, the anonymization described above, and controlled access for registered teams, so that the diagnostic benefit of quantifying identifiability outweighs the additional risk. Please contact us if you are interested in participating.

Acknowledgments

Funded by grants from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project numbers 539645652 and 539634240 – within the priority programme "New Data Spaces for the Social Sciences" (SPP 2431).

References

Phillip L. Ackerman. 1996. A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, 22(2):227–257.

Phillip L. Ackerman. 2000. Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 55(2):P69–P84.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: a survey of unlearning methods. *Artificial Intelligence Review*, 58(3). Article 90.

Raymond B. Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1):1–22.

Albert T. Corbett and John R. Anderson. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Clemens Deußer, Steffen Passmann, and Thorsten Strufe. 2020. Browsing unicity: On the limits of anonymizing web tracking data. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 777–790.

Cody V. Dong, Qihong Lu, Kenneth A. Norman, and Sebastian Michelmann. 2025. Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*, 29(10):928–941.

Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 72–79, New York, NY, USA. Association for Computing Machinery.

Johannes C. Eichstaedt, Margaret L. Kern, David B. Yaden, H. Andrew Schwartz, Salvatore Giorgi, Gregory Park, Courtney A. Hagan, Victoria A. Tobolsky, Laura K. Smith, Anneke Buffone, Jonathan Iwry, Martin E. P. Seligman, and Lyle H. Ungar. 2021. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4):398–427.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W. Domingue, Emma Brunskill, and Noah D. Goodman. 2024. Psychometric alignment: Capturing human knowledge distributions via language models. *Preprint*, arXiv:2407.15645.

Markus J. Hofmann, Markus T. Jansen, Christoph Wigbels, Benny Briesemeister, and Arthur M. Jacobs. 2024. Individual text corpora predict openness, interests, knowledge and level of education. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (CogALex) @ LREC-COLING 2024*, pages 14–25, Torino, Italia. ELRA and ICCL.

Markus J. Hofmann, Lara Müller, Andre Rölke, Ralph Radach, and Chris Biemann. 2020. Individual corpora predict fast memory retrieval during reading. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 1–11, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Gary Marchionini. 2006. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46.

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT-2. https://github.com/bminixhofer/gerpt2.

Amelie Nikstat, Angelina Höft, Jessica Lehnhardt, Stephanie Hofmann, and Christian Kandler. 2018. Entwicklung und Validierung einer Kurzversion des Fragebogeninventars für Freizeitinteressen (FIFI-K). *Diagnostica*, 64(1):14–25.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Qdrant Team. n.d. Qdrant: Vector Search Engine. https://qdrant.tech/. Accessed 2026-04-16.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Stefan Schipolowski, Oliver Wilhelm, Ulrich Schroeders, Anastassiya Kovaleva, Christoph J. Kemper, and Beatrice Rammstedt. 2013. BEFKI GC-K: A short scale for the measurement of crystallized intelligence. *Methods, Data, Analyses*, 7(2):153–181.

Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778.

Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. 2007. Information re-retrieval: Repeat queries in Yahoo's logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 151–158. ACM.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2025. Personalization of large language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. 2025. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13391, Vienna, Austria. Association for Computational Linguistics.

Luc Zimny, Ulrich Schroeders, and Oliver Wilhelm. 2024. Ant colony optimization for parallel test assembly. *Behavior Research Methods*, 56:5834–5848.

A Example Items

The 129 fine-tuning items and the 22 held-out selection items were drawn from the openly available general-knowledge item pool of Zimny et al. (2024).[4] The following item (`mus073`, solution rate .49 in the pool) illustrates the format. All items were administered in German; the English translation in brackets is taken from the pool documentation, and the correct option is printed in bold.

> Welches Instrument gehört nicht zu der Gruppe der Holzblasinstrumente?
> [Which instrument does not belong to the group of woodwind instruments?]
>
> ☐ **Waldhorn** [**french horn**]
> ☐ Querflöte [transverse flute]
> ☐ Oboe [oboe]
> ☐ Fagott [bassoon]

The 12 BEFKI GC-K core items follow the same four-option format and are documented in Schipolowski et al. (2013), whereas the 24 extended items are deliberately not published, to keep them out of public text collections as the contamination control, in line with standard practice for

[4] https://osf.io/u68nk/

psychological test material. The following core item, reproduced from the public BEFKI GC-K documentation in the GESIS instrument archive ZIS[5], illustrates the format; as no official English translation exists, the bracketed translation is our own.

> Wozu dient die Mitose?
> [What is the function of mitosis?]
>
> ☐ Stoffwechselregulation [metabolic regulation]
> ☐ Fortpflanzung [reproduction]
> ☐ Bildung von Keimzellen [formation of germ cells]
> ☐ **Zellvermehrung bei Wachstumsvorgängen [cell proliferation during growth]**

B Comparison of Further Models

To probe whether the calibration gap identified in the main analysis is tied to model scale, we evaluated a set of additional models under the identical IC-based RAG pipeline, retrieval configuration, and probability extraction (Table 3). While larger models tend to show more correct answers, the fine-grained log-loss metric indicates that smaller models should be more suitable for individualized knowledge simulation.

Model	Params	Corr.	MA	LL:M	LL:C	CK
Qwen3-0.6B[a]	0.6B	.46	.33	2.24	1.82	.43
LFM2-700M	0.7B	.65	.52	2.22	1.10	.62
Qwen3-1.7B[a]	1.7B	.67	.31	6.29	1.56	.54
Qwen3-4B-Instruct-2507	4B	.83	.57	5.47	1.07	.60
Mistral-7B-v0.3	7B	.88	.60	3.07	0.74	.63

Table 3: Model Comparison Under the Identical IC-Based RAG Pipeline

Note. Corr. = Correctness; MA = Match accuracy; LL:M = Log loss: match; LL:C = Log loss: correct; CK = CK-accuracy. All rows except the Qwen models are unadapted base checkpoints evaluated under the identical retrieval configuration and probability extraction, without task-specific fine-tuning; $N = 484$–498; values are means over participant×item pairs in the with-context condition. Chance level: .25 (MA), 1.39 nats (LL); majority-class baseline for CK: .66. [a]Fine-tuned on the QA task ($N = 316$).

Across the tested models the individual context barely shifts answer behavior: for Mistral-7B, correctness (.875 vs. .861) and match accuracy (.596 vs. .609) are near-identical with and without retrieval, and the shift is smallest for the model with the strongest parametric knowledge. This is expected if the answer distribution is dominated by a sharp parametric prior that the retrieved context cannot outweigh. Our target, however, is not correctness but accuracy in the sense of reproducing a participant's answer behavior, and a strong prior maximises the former while leaving little room for the latter. The bottleneck is therefore not model size but how the individual signal enters the model: retrieval adds it as inference-time evidence that competes with the prior and loses. A consequent next step, sketched in the Conclusions, is to write each participant's corpus into the model weights, where the prior itself resides, with per-participant adaptation of smaller models as a secondary lever for a less dominant prior.

[5] https://doi.org/10.6102/zis220