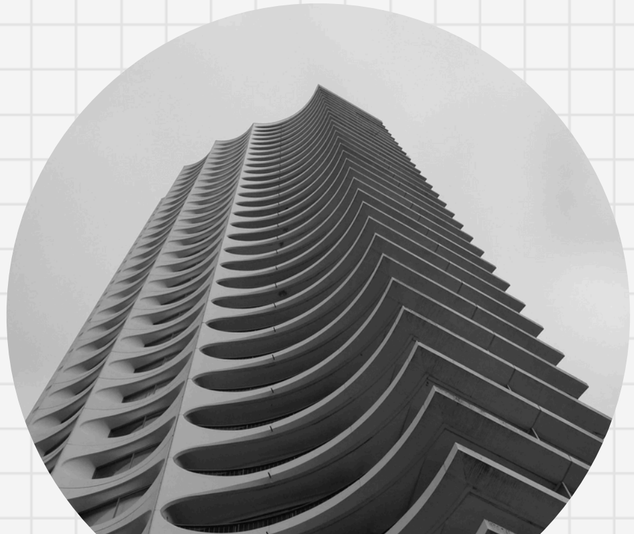
PERSONALIZED INTERVIEW COACHING VIA DIGITAL TWINS AND RAG

Cynthia Fu, Eric Huang, Yurun Guo,
Yi-Ting Cheng, Chao Wang
12/17/2025



Agenda

1. Motivation and Use Case
2. System overview & Data Strategy
3. Data Source & Processing
4. RAG and Prompt engineering
5. Inference Output
6. Model Evaluation

Project Background- Motivation & Problem

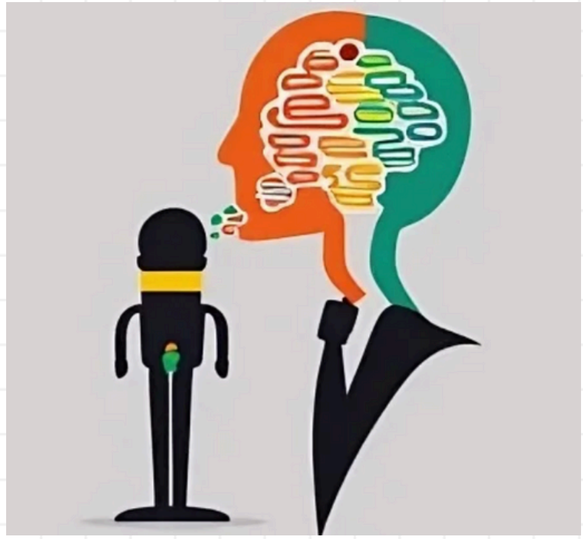
Why Interview Prep Needs a Digital Twin

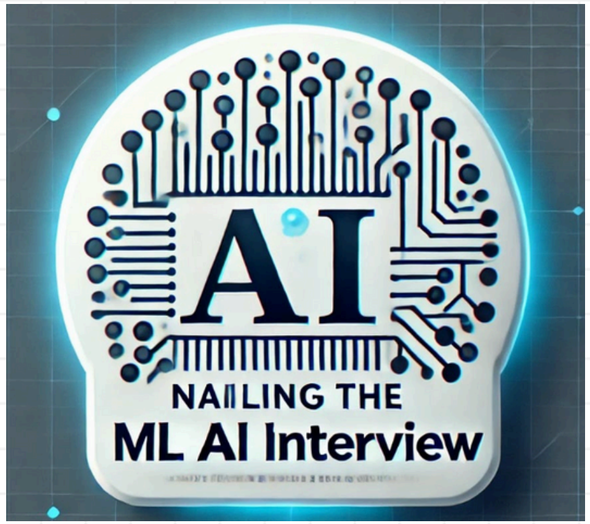
- Interview preparation for Data Science / Data Analyst roles is largely generic and one-size-fits-all
- Existing resources (blogs, videos, question lists) do not adapt to: Candidate background/Skill gaps/Time constraints and role-specific expectations

- Candidates over-prepare low-impact topics
- Under-prepare critical weaknesses
- Spend significant time without measurable progress

Key Gap

- Expert-level interview coaching exists, but is:
  - Expensive
  - Not scalable
  - Not personalized beyond surface-level advice





Project Background- Use Case, User, and Business Value

Target User

- **Interns, New graduates**
- **Early-career candidates,**
- **Junior professionals preparing for role transitions**

Proposed Solution

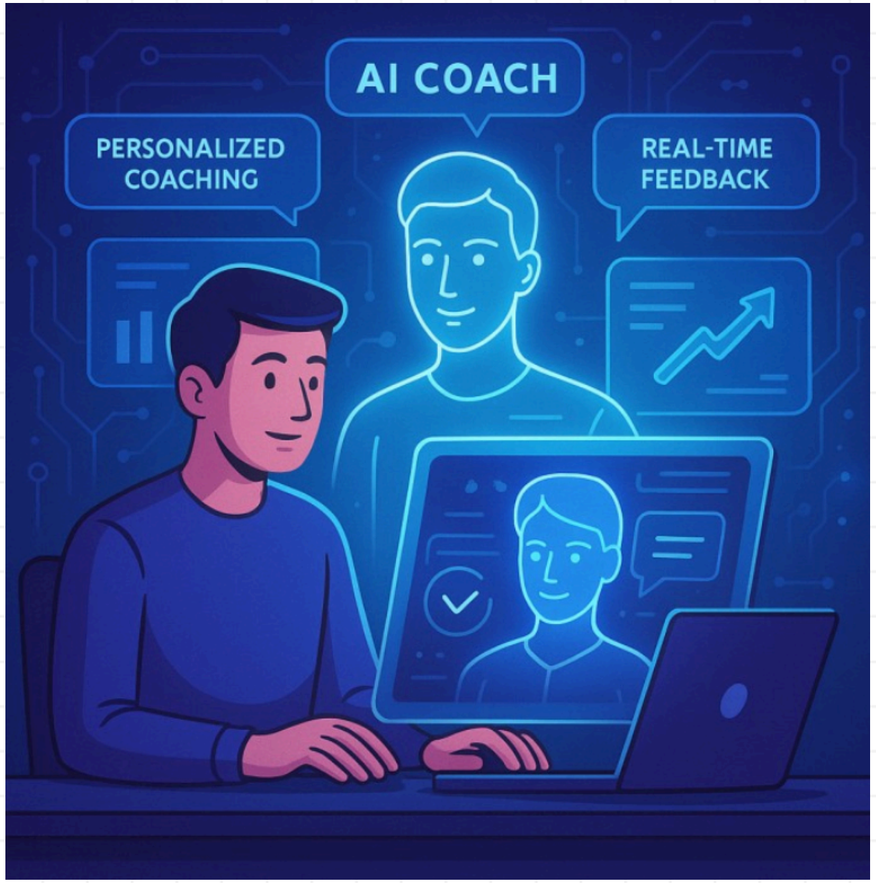
- A Digital Twin ofthecandidate’sbackground
- Combines: Role-based persona design (Digital Twin), Curated interview knowledge (YouTube + GitHub), RAG
- Generates: Personalized interview preparation plans, Prioritized topic recommendations, Actionable guidance aligned with real interview expectations

Business Value

- Scales expert coaching at near-zero marginal cost
- Improves preparation efficiency by:
- Focusing on high-impact gaps, Reducing redundant or misaligned study

Applicable to:

- Career coaching platforms, University career services



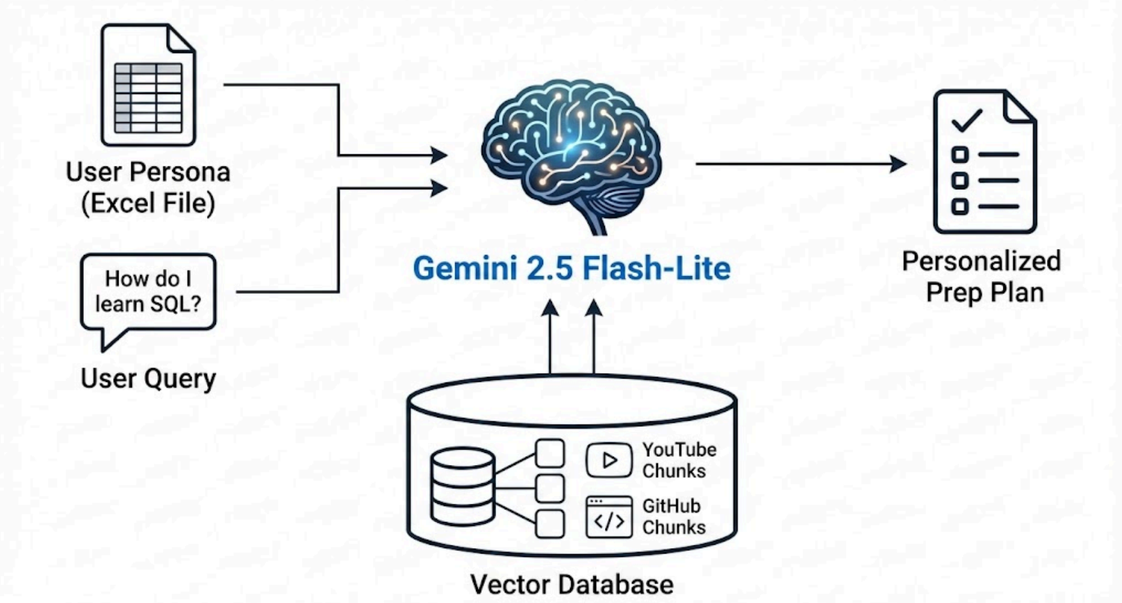
System Architecture: The "Digital Twin" Career Coach

Core Concept:
A Retrieval-Augmented Generation (RAG) system acting as a personalized Data Science Interview Coach

The "Brain" (LLM):
Powered by Google Gemini 2.5 Flash-Lite

Key Components:

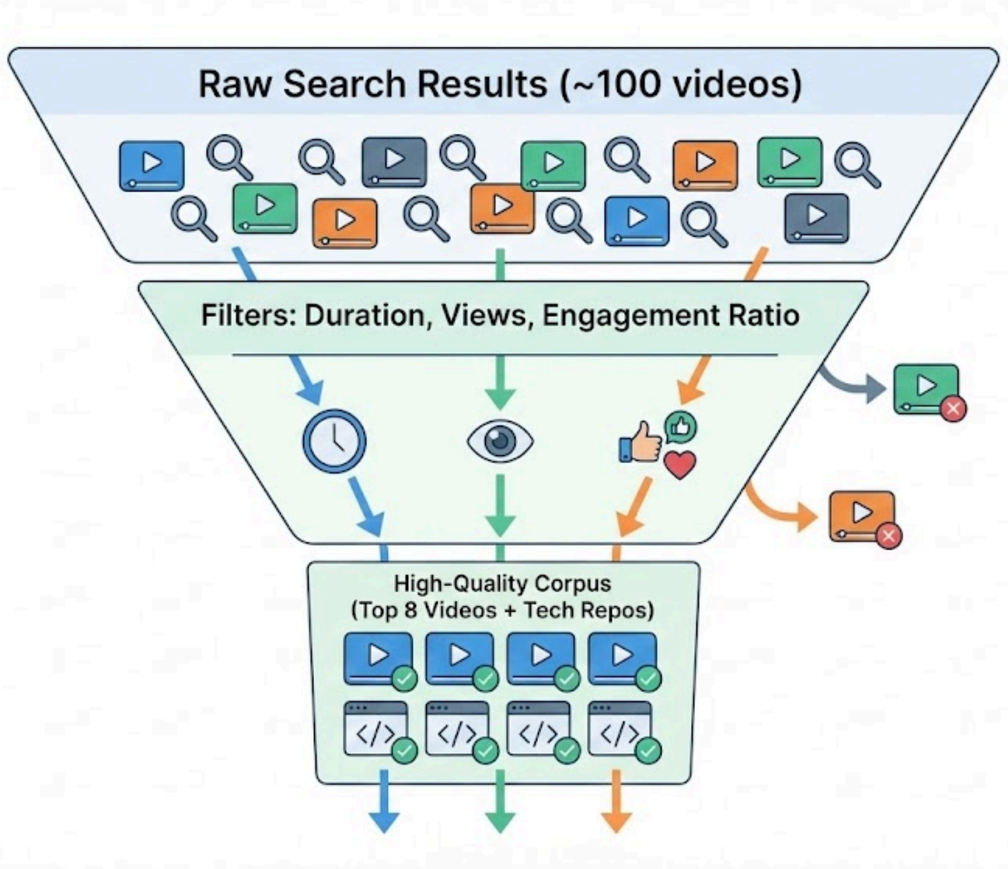
- **Orchestrator:** Manages the flow between user personas, retrieval, and generation.
- **Vector Store:** Uses models/text-embedding-004 to index knowledge.
- **Retrieval:** Semantic search (Cosine Similarity) fetches top-k relevant advice chunks.
- **Persona Injection:** Dynamically adapts answers based on user profiles (e.g., "Junior DS" vs. "Senior Analyst") loaded from role_based.xlsx.



Data Strategy: Curating High-Quality Career Advice

Primary Source: YouTube (Conversational Strategy)

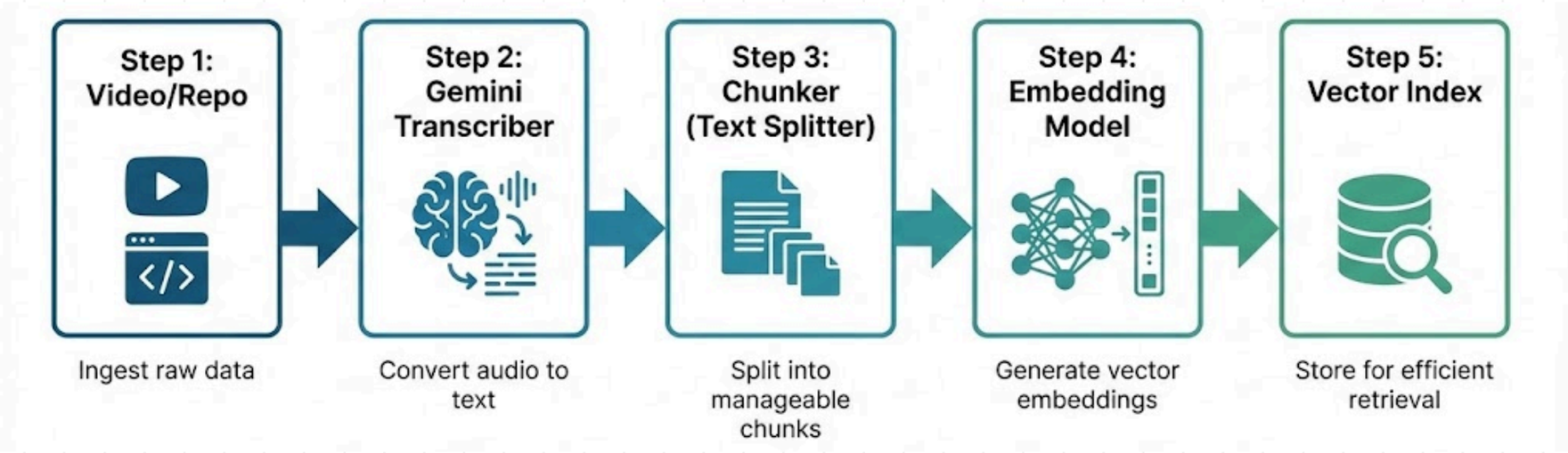
- **Automated Quality Filters:**
  - **Authority:** >10k Views
  - **Relevance:** 2–45 min duration (No Shorts, no 3-hour streams)
  - **Quality:** High Like/View ratio (>2%) to ensure community validation
  - **Logic:** 1.5x score boost for titles containing "Roadmap" or "Plan"



Secondary Source: GitHub (Technical Depth)

- **Curated technical interview repositories** (e.g., Devinterview-io, khangich)
- **Provides precise definitions (SQL syntax, ML theory)** to complement YouTube's broad advice

Data Preprocessing: From Raw Content to Knowledge



Ingestion

- **YouTube:** yt-dlp downloader (handling cookies/blocks)
- **Transcription:** Gemini 2.5 Flash converts audio to text
- **GitHub:** Raw Markdown ingestion and cleaning

Chunking Strategy

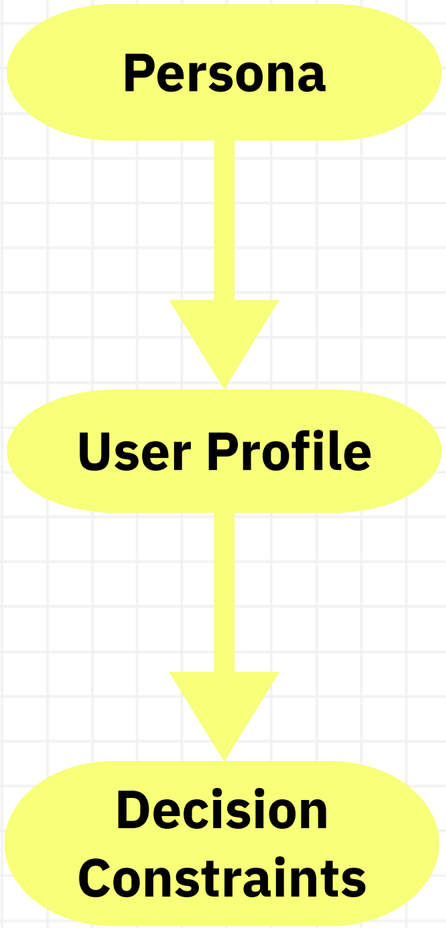
- **Splits long transcripts into 500-word chunks with 50-word overlap**
- **Why?** Allows the system to retrieve **specific advice** (e.g., "Week 2 SQL Plan") rather than entire transcript

Embedding

- Model: **models/text-embedding-004**
- Converts text chunks into 768-dimensional vectors for semantic search

Prompt Engineering: Encoding User Preferences

Prompt = Persona + User Profile + Decision Constraints



How user preferences are used

- User preferences are explicitly injected into the system prompt
- Preferences include:
  - Target role (Junior DS, Entry-level DS, Junior DA)
  - Preparation timeline (e.g., 4 weeks, 10 days)
  - Skill gaps (e.g., weak SQL, limited experiment design)
- Preferences are loaded from structured persona profiles

Prompt Engineering Iterations: From Generic LLM to Interview Coach

Baseline Prompt

- Generic assistant
- No persona or constraints
- High-level, generic advice

Persona-only Prompt

- Interview coach persona added
- Better tone and structure
- Still lacks prioritization

Final Prompt (Persona + Constraints)

- Step-by-step preparation plans
- Role- and time-aware prioritization
- Consistent, actionable outputs

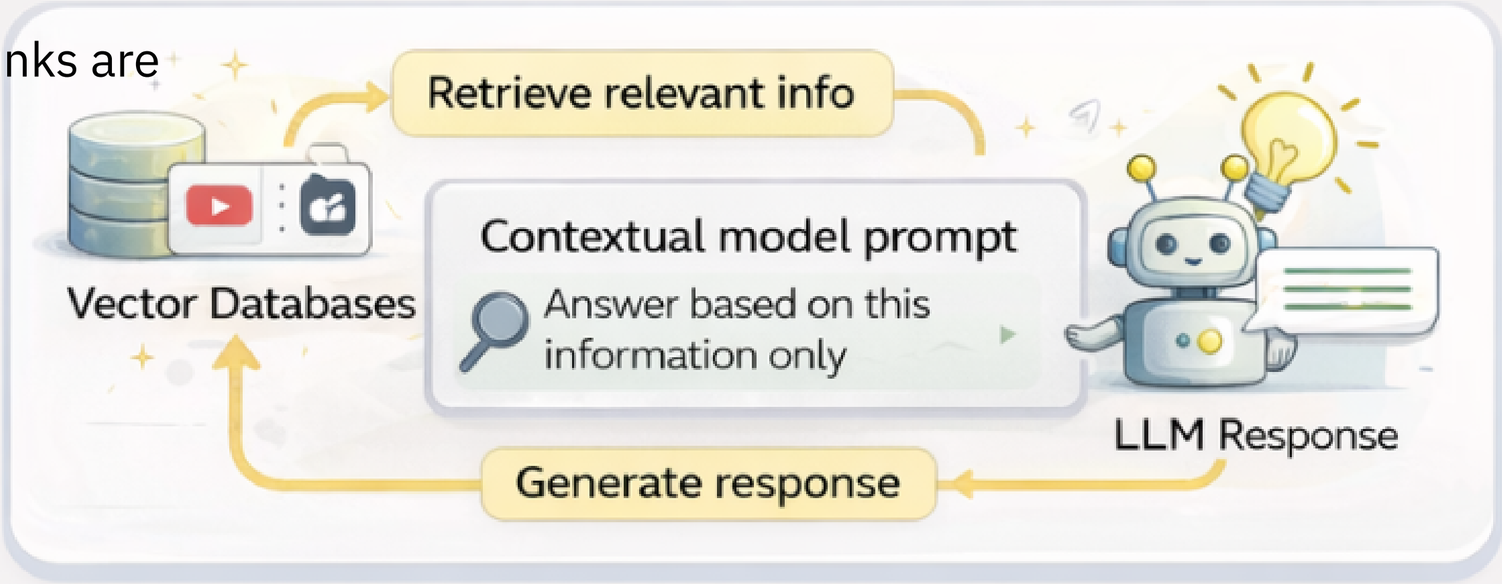
Retrieval-Augmented Generation (RAG): Grounding Expert Advice

What RAG does in our system

- Expert interview advice collected from YouTube and GitHub
- Content split into semantically meaningful chunks
- Chunks embedded and indexed in a vector database
- For each query, top-k relevant chunks are retrieved

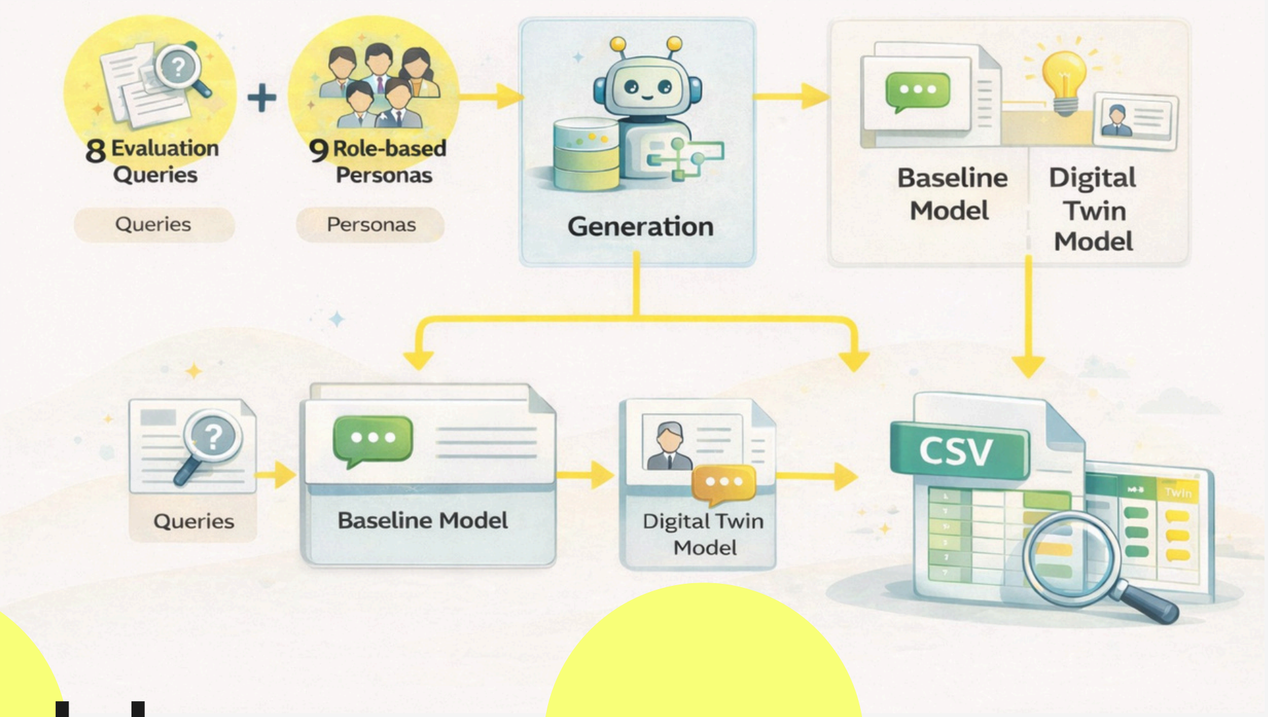
Key design choice

- Retrieved expert advice is explicitly injected into the prompt
- Model is instructed to answer based only on retrieved context
- Prevents hallucination and generic responses



Prompt engineering improves how the advice is delivered, while RAG improves what the advice is grounded in.

Inference Outputs for Evaluation



8 queries 9 role-based personas

- Fixed set of 8 evaluation queries covering common DS/DA interview scenarios.
- 9 role-based personas used to condition the digital twin.

2 Models

For each (role, query) pair, we generate:

- Baseline response
- Digital Twin response (RAG + persona)

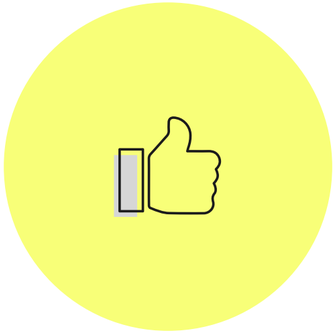
9 × 8 roles/queries

- Final output saved as a single CSV file for consistent manual evaluation
- Total inference outputs: 9 × 8 roles/queries
- 72 rows, baseline + twin per row

Query: I struggle with experiment design and A/B testing interview questions.

Baseline (Generic Response)

- Lists books, blog posts, and example interview questions
- Focuses on external resources rather than a preparation strategy
- Lacks prioritization or a clear study plan



Digital Twin (RAG + Persona)

- Identifies key A/B testing concepts to prioritize
- Proposes a clear 4-week preparation plan
- Anticipates common pitfalls and interview expectations

Example Output Comparison: Baseline vs Digital Twin

Key Resources:

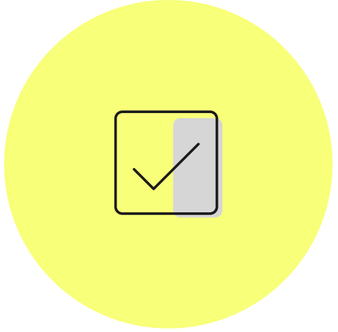
* **"Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing"**: This is a highly recommended
* **LinkedIn Engineering Blog**:
  * "Detecting Interference: An A/B Test of A/B Tests" (Found in appliedml.md)
  * "Our evolution towards T-REX: The prehistory of experimentation infrastructure at LinkedIn" (Found in appliedr
  * "Building Inclusive Products Through A/B Testing" (Found in appliedml.md)

Practice Questions & Topics to Prioritize:

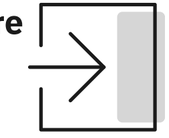
* **Designing A/B Tests:**
  * "How would you design an A/B test for the homepage?" (Found in questions.md)
  * "We have a product that is getting used differently by two different groups. What is your hypothesis about why
  * "Explain how network effects might influence your choice of how to assign experimental/control units and mea
* **Understanding Interference:** The concept of "interference" in A/B tests is important, as highlighted by the Link
* **Practical Applications:** Questions about how to use A/B testing for specific scenarios, like building inclusive p

General Advice for Answering Technical Questions (from faqs.md):

* "Think about interview's intentions then answer to the point. Try to stay precise, concise and to the point."



Under the same query, the digital twin provides more structured and personalized guidance compared to the baseline.



Alright, let's get down to business. You've got a tight deadline, and we need to be incredibly focused. Your goal is to nail those A/B testing and experiment design questions, and with four

Priorities

Given your background and timeline, we need to laser-focus on the core concepts of A/B testing and experiment design. Anything that isn't directly supporting these areas will have to wai

1. **Core A/B Testing Concepts:** Understand the fundamental principles. This includes defining metrics, hypothesis formulation, sample size calculation (even conceptually), and interpre
2. **Experiment Design Principles:** This is where your weakness lies, so we'll attack it head-on. Think about randomization, control vs. treatment groups, and potential biases.
3. **Common Pitfalls in A/B Testing:** Learn what can go wrong and how to identify and mitigate these issues.

Weekly Plan

This is a sprint. Every week builds on the last.

* **Week 1: Foundations of A/B Testing**
  * **Focus:** Grasp the "why" and "what" of A/B testing.
  * **Topics:** What is A/B testing? What kind of problems does it solve? What are key metrics? How do you form a hypothesis?
  * **Action:** Read the "Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing" (mentioned in the `README.md` context). It's your primary text for this week.

* **Week 2: Designing Experiments**
  * **Focus:** Dive into the mechanics of setting up an experiment.
  * **Topics:** Randomization, defining user segments, choosing experiment units (e.g., user vs. session), and considerations for different platforms (e.g., web vs. app).
  * **Action:** Study LinkedIn's blog posts on A/B testing (e.g., "Detecting Interference: An A/B Test of A/B Tests", "Building Inclusive Products Through A/B Testing"). These will give you

* **Week 3: Interpretation and Pitfalls**
  * **Focus:** Understanding results and common errors.
  * **Topics:** Statistical significance (conceptual understanding is key here, not deep theory), p-values, confidence intervals, and common errors like Simpson's Paradox, novelty effects
  * **Action:** Review the LinkedIn articles again, looking specifically for how they discuss results and potential issues. Tackle practice questions that focus on interpretation.

* **Week 4: Practice and Refinement**
  * **Focus:** Applying knowledge and simulating interview scenarios.
  * **Topics:** All previous topics, with an emphasis on answering common A/B testing interview questions.
  * **Action:** Work through A/B testing specific questions from the `questions.md` context. Simulate answering them out loud.

Practice Resources

You need to be strategic with your time.

1. ***Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing***: This is your bible. Read it cover-to-cover, focusing on the practical aspects.
2. **LinkedIn Blog Posts**:
   * "Detecting Interference: An A/B Test of A/B Tests"
   * "Building Inclusive Products Through A/B Testing"
   These offer real-world examples and practical considerations.
3. **`questions.md` Context**:
   * **Question 76: "LinkedIn question – How would you design an A/B test for the homepage?"** - This is gold. Practice this one extensively. Think about what metrics you'd track, what
   * **Question 64: "Netflix question – When you split a population for A/B testing, what are some reasons you could see a significant difference in the control and var"** - This probes you
   * **Question 81: "We have a product that is getting used differently by two different groups. What is your hypothesis about why and how would you go about testing it?"** - This tests y
4. **Quora / Blog Posts**: Look for articles on "common A/B testing interview questions" if you need more practice scenarios, but prioritize the provided resources first.

Pitfalls to Avoid

* **Getting Lost in Statistics:** You're not aiming to be a statistician, but a practical data scientist who can *apply* statistical thinking. Focus on the implications of statistical concepts for
* **Overcomplicating Designs:** For a 4-week sprint, stick to fundamental, robust A/B test designs. Don't get bogged down in complex multi-variate tests unless the question specifically
* **Failing to Define Success Metrics:** This is a common mistake. Always, *always* define your primary and secondary metrics clearly before discussing any experiment.
* **Ignoring Potential Biases/Interference:** Always think about what could go wrong and how you'd mitigate it. The LinkedIn article on interference is a great starting point.
* **Not Practicing Out Loud:** You need to be able to articulate your thoughts clearly and concisely under pressure. Practice answering these questions as if you were in the interview.

This is a challenging timeline, but with laser focus on these areas, you can significantly improve your performance in A/B testing and experiment design questions. Let's get to work!

Evaluation Design & Setup

Models Compared

- Baseline (Zero-Shot RAG): Retrieved context is directly injected into the prompt
- Digital Twin (Advanced Prompting + RAG): adds role prompting, persona-based constraint (role, timeline, weaknesses)

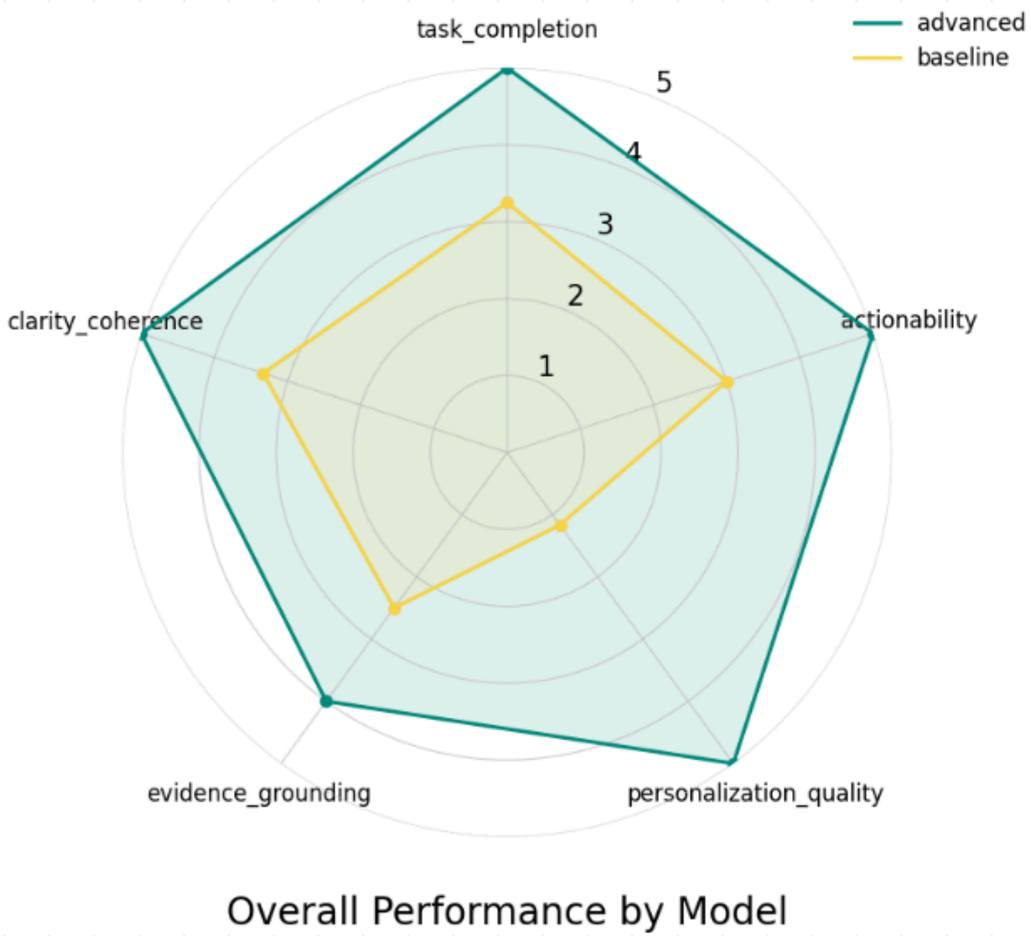
Evaluation Method & Rubrics

- Manual evaluation using 1–5 scale, with scores averaged across five rubric dimensions
- 5 Metrics: Task Completion, Actionability, Personalization Quality, Evidence Grounding, and Clarity & Coherence

Test Set & Sampling

- Roles: Junior Data Scientist (4 weeks), Entry-level Data Scientist (10 days), Junior Data Analyst (1.5 months)
- Queries: Interview planning, SQL practice, A-B testing, and behavioral (STAR)

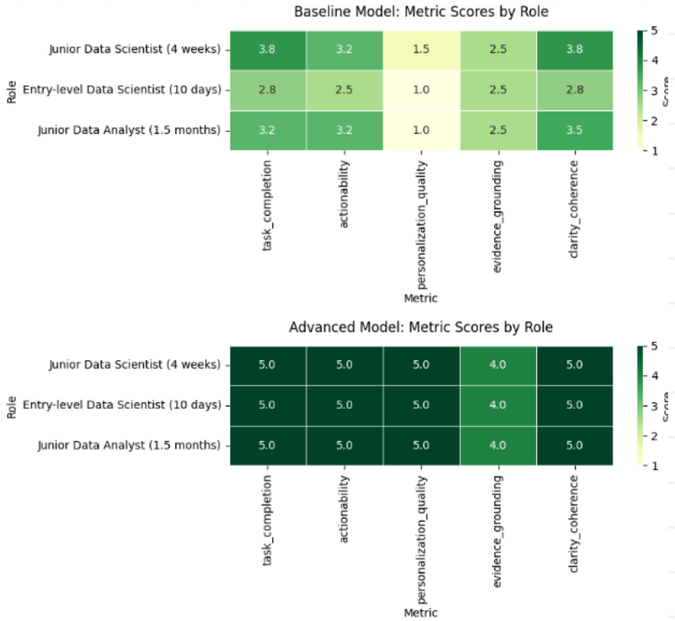
Results



Overall Performance by Model

Key Results

- Digital Twin outperforms baseline across all metrics
- Largest gains in: Personalization Quality, Actionability, Clarity & Coherence
- Baseline performance varies by role and query
- Digital Twin remains consistently high across roles and queries



What Improved

- Role-aware prioritization under time constraints
- Concrete, step-by-step guidance
- Reduced variance across users and questions
- Shift from generic advice → interview coaching

What Improved Less

- Evidence grounding capped by retrieval quality
- Score saturation near upper bound (4–5)
- Manual evaluation does not scale

Business Impact & Limitations

Business Impact

- **Prompt design clearly matters**: The Digital Twin outperforms the baseline using the same model and retrieval, showing gains come from prompting, not model changes
- **Personalization actually works**: Injecting role, timeline, and weaknesses produces clearer and more actionable guidance
- **Scales without retraining**: High quality coaching is achieved without fine tuning, enabling broad role coverage at low cost
- **Product ready direction**: Supports AI interview coaching, role specific prep tracks, and adaptive learning plans

Limitations & Future Work

- **Limited evaluation scope**: Current test set covers a small number of roles and queries, which may not fully represent the diversity of real world interview scenarios
- **Human scored evaluation bias**: Manual scoring introduces subjectivity and inter rater variance despite rubric standardization
- **Retrieval quality not deeply explored**: The study isolates prompt effects but does not optimize or vary retrieval depth, document diversity, or ranking strategies

Thank you for listening!