An Analysis of LLM-Driven Semantic Matching Frameworks for Complex Domains

Executive Summary: The New Architecture for Semantic Matching

This report provides a comprehensive technical and strategic analysis of the current state of algorithmic matching, critically evaluating a proposed Large Language Model (LLM)-driven framework. The central thesis is that traditional "match score" systems are obsolete. The foundational algorithms in modern dating, such as Collaborative Filtering, and in recruitment, such as keyword-based Applicant Tracking Systems (ATS), are demonstrably flawed. They suffer from systemic bias amplification, a fundamental lack of semantic understanding, and a critical misalignment of incentives that favors platform engagement over user success.

The proposed three-stage, LLM-driven architecture—comprising (1) Semantic Profile Generation, (2) Pairwise Trade-off Analysis, and (3) Personalized Semantic Filtering—is not merely hypothetical. It represents the emergent, state-of-the-art paradigm for high-stakes, nuanced matching. Research from 2024-2025 confirms this architectural pattern is being actively deployed and researched in fields ranging from precision medicine to legal technology.

However, a naive implementation of this framework presents technical, ethical, and financial challenges that are catastrophic in scale. This report finds three critical barriers:

1. **Computational Intractability:** The proposal for an exhaustive, pairwise comparison ("for each I and each j") is a computationally intractable O(n^2) problem. This "quadratic complexity" would be financially ruinous and operationally non-viable.
2. **Algorithmic Bias:** General-purpose LLMs, far from solving bias, introduce new and more insidious forms. Research demonstrates that off-the-shelf foundational models exhibit significant, systemic intersectional racial and gender biases in hiring, demonstrably performing *worse* than existing systems.
3. **Data Privacy and Security:** The framework's reliance on consuming vast quantities of unstructured, sensitive personal data (e.g., medical records, private chats, full resumes) creates profound privacy, security, and regulatory compliance risks.

The framework's viability is therefore entirely dependent on three strategic pivots that address these challenges. The analysis concludes that the correct implementation must move:

- From **Exhaustive Comparison** to a **Hybrid, Multi-Stage "Retrieve-and-Rerank"** architecture.
- From **General-Purpose LLMs** to **Audited, Domain-Specific, Fine-Tuned Models**, which are proven to be both more accurate *and* more fair.
- From **Public API Calls** to a **Private-Cloud or Locally-Deployed** architecture to ensure "privacy-by-design".

Ultimately, this report validates the proposed architecture's conceptual soundness while providing a rigorous, evidence-based roadmap for navigating its significant implementation hurdles.

The State of Algorithmic Matching (2024-2025): A System of Scores and Filters

The premise that we are "past the point of match scores" is predicated on the well-documented failures of current-generation matching systems. An analysis of the two primary domains, dating and recruitment, reveals systems that are not only semantically weak but are often optimized for goals contrary to the user's, such as revenue and filtering efficiency, rather than optimal matching.

A. Analysis of the Dating Market: Elo, Bias, and Homogeneity

Modern dating applications employ a layered stack of algorithms, but the most dominant and consequential are not necessarily the most sophisticated.

- **Core Mechanisms:** The primary filters are user-set preferences (age, gender) and geographic proximity. Beyond this, platforms employ two main algorithmic drivers:
    1. **Desirability Scores:** Tinder, for example, utilizes an "Elo rating," a score based on a user's swipe behavior and the behavior of those who swipe on them. This score dictates whose profiles are shown and in what order.
    2. **Collaborative Filtering (CF):** This is the most common machine-learning approach, powering recommendations on platforms from Hinge to Amazon. CF operates on the principle of "users who liked X also liked Y". If User A and User B both swipe right on the same three profiles, the system infers they have similar tastes and will begin recommending other profiles that User B liked to User A. Some platforms, like OkCupid, also use content-based filtering, generating a "Match % score" from user-answered questions about preferences and "deal breakers".
- **Critical Trade-offs and Failures:**
    - **Misaligned Incentives:** The "Elo score" system reveals a fundamental, structural misalignment of goals. The algorithm is not optimized for user compatibility or relationship success; it is optimized for platform engagement and revenue. The "desirability score" is used to "manipulate match visibility" and "create a sense of artificial scarcity". This manufactured scarcity "drives urgency—and upgrades," nudging users toward paid subscriptions. This incentive structure is inherently at odds with finding a user the "perfect match," as a successful match results in a lost user and revenue.
    - **Collaborative Filtering as a Bias Amplifier:** The core trade-off of Collaborative Filtering in a social domain is that it functions as a powerful *bias and homogeneity engine*. CF algorithms do not understand *why* users make choices; they only observe the choices themselves. Because users' "revealed preferences" (swipes) often contain deep, implicit racial and physical biases, the CF algorithm *learns, codifies, and scales* these biases. The system doesn't just reflect a biased world; it *enforces* it by "deepening existing racial biases" and "homogenizing behavior". This leads to a feedback loop where users are shown an increasingly narrow, homogenous set of profiles, directly contradicting the goal of novel, compatible discovery.

B. Analysis of the Recruitment Market: The Tyranny of the Keyword

The recruitment market is dominated by a different, but equally flawed, technology: the Applicant Tracking System (ATS). An estimated 99% of Fortune 500 companies rely on an ATS to manage hiring.

- **Core Mechanisms:** The ATS is, first and foremost, a filtering and database management tool. Its primary mechanism is *resume parsing* based on rigid *keyword matching*. The system "reads" a resume, extracts text, and compares the frequency and presence of keywords (e.g., "Python," "SQL," "project management") against the keywords in the job description. It then assigns a "resume score" or "match percentage" to rank candidates. Recruiters, facing hundreds of applications, often only review the top-scoring results.
- **Critical Trade-offs and Failures:**
    - **The "False Negative" Catastrophe:** The keyword-matching paradigm is "fundamentally flawed". Its defining trade-off is an astronomically high "false negative" rate—the rejection of qualified candidates. Some estimates suggest these systems screen out *up to 75%* of applicants due to rigid rules and poor font recognition. Reinforcing this, data indicates *88% of employers* believe they are losing qualified candidates because their resumes are not "ATS-friendly".
    - **The Semantic Failure:** The system fails because it is lexical, not semantic. It lacks *contextual understanding*. The system cannot comprehend synonyms, equivalent experiences, or nuanced job titles. A widely cited example is a candidate with the title "Product Lead" at a major company being automatically rejected for a "Product Manager" role, despite having identical qualifications. Similarly, a "Software Engineer II" might be missed for a "Backend Developer" role, even if the underlying skills are a perfect match. The ATS has no concept of "equivalency of experiences". * **The "Keyword Optimization Death Spiral":** The failure of the ATS has created a counter-productive and absurd "arms race". Candidates, aware of these flawed systems, now use AI tools to *optimize CVs with perfect keywords*. In response, companies deploy more AI to *filter these AI-generated resumes*. The result is an "endless arms race where keywords become meaningless". This death spiral is the ultimate evidence that the keyword-matching paradigm is obsolete. The "match" must be escalated from the lexical level to the semantic, or LLM, level.

This analysis of the dominant matching systems in dating and hiring reveals a clear and urgent need for a new architecture.

Table 1: Comparative Analysis of Current Matching Algorithms (Dating vs. Jobs)

Domain	System Example	Core Mechanism	Key Trade-off / "The Problem"
Dating	Tinder	**Desirability Score (Elo)**	**Misaligned Incentives:** Score is gamed to "manipulate match visibility" and "create artificial scarcity" to drive revenue, not to find the best compatible match.
Dating	Hinge, Bumble	**Collaborative Filtering**	**Bias Amplification &**

Domain	System Example	Core Mechanism	Key Trade-off / "The Problem"
		(CF)	**Homogeneity:** Learns and exacerbates existing user biases (e.g., racial). Leads to "homogenization of behavior" rather than novel, compatible discovery.
Jobs	Workday, Taleo	**Applicant Tracking System (ATS)**	**Semantic Rigidity & False Negatives:** "Fundamentally flawed" rigid keyword matching screens out *up to 75%* of applicants and *88%* of employers believe they lose qualified candidates. Fails to equate "Product Lead" with "Product Manager".

The User's Proposition: An Analysis of an LLM-Driven, Trade-off-Based Matching Framework

The proposed three-stage framework addresses the failures of current systems by shifting the paradigm from *scores* to *semantics* and from *gatekeeping* to *explainability*. This model is strongly aligned with state-of-the-art research in personalization and AI.

A. Component 1: LLM-Based Profile Generation

The first component, "LLM driven matches can consume match subject i1…in and match subject j1…jn build profiles based on plain English," describes a sophisticated method of user modeling that leverages the unique capabilities of LLMs to understand unstructured data.

- **Technical Validation:** This is a significant evolution. Traditional systems rely on sparse interaction matrices (for CF) or hand-engineered features. The 2025 "PURE" framework details an LLM-based system that can *build and maintain evolving user profiles* by systematically extracting "likes," "dislikes," and "key features" from unstructured user reviews. Similarly, the "LLM-TUP" model generates *natural language representations* of user histories to model long-term and short-term preferences. demonstrates this by generating "interpretable natural language user profiles" from millions of tweets.
- **Enterprise Application:** This is not limited to users. A 2025 report from DoorDash details their strategic shift *away* from opaque embeddings *toward* "rich, narrative-style profiles written in natural language" for all their core entities: consumers, merchants, and items. This allows them to capture semantic nuance impossible for traditional systems, such as "prefers spicy Sichuan dishes, avoids dairy".

- **The "Interpretable" and "Editable" Profile:** The key advantage of this component, as identified by DoorDash , is that these LLM-generated profiles are *human-readable, interpretable, and editable*. An opaque embedding vector (e.g., "cosine similarity 0.83") is a black box. A profile that reads "prefers spicy food" is transparent and can be *corrected by the user in plain English*. This is a paradigm shift in user modeling, enabling a new level of accuracy and user control.

B. Component 2: The "Match Trade-Off" Engine

The second component, "consider match(in, jn) for each I and each j, output match_trade offs for each match," forms the innovative core of the proposal. It replaces the information-poor "match score" with an information-rich, explainable report. This is strongly validated by emerging research in explainable AI (XAI) and "LLM-as-a-Judge."

- **In Recruitment:** This is actively being developed. A 2025 paper on a multi-agent framework for hiring describes a "summarizer agent" that generates a report highlighting "key strengths and pinpointing missing competencies" in bullet points. This allows recruiters to "efficiently compare multiple candidates". describes a similar "resume summarizer" that generates a "concise, easy-to-understand report, highlighting the candidate's strengths and areas for improvement." LLMs can interpret resumes to extract not just *explicit skills* (e.g., SQL, Python) but also *implicit concepts* (e.g., "data-driven decision making") that are semantically aligned with the job description. This provides recruiters with a nuanced "pro/con" analysis rather than a simple "pass/fail" score.
- **In Recommendations:** In media, this is known as *explainable recommendation*. A 2025 user study on movie recommendations *proves* this concept's value. It found that "contextualized explanations" (i.e., *why* a match is good, "because you liked X") are highly effective. These "trade-off" reports "effectively meet users' cognitive needs" (fostering trust and transparency) and significantly "increas[e] users' intentions to watch recommended movies".
- **The AI as "Decision-Support Co-pilot":** This component fundamentally reframes the role of AI in high-stakes decisions. The "match score" in a traditional ATS (Section II.B) acts as a *gatekeeper*—it makes a decision *for* the human, filtering out 75% of applicants. The "match trade-off" report acts as a *co-pilot*—it provides synthesized intelligence *to* the human, empowering *them* to make a better, more informed decision.

C. Component 3: The "Personalized Trade-Off Space"

The third component, "apply LLMs to extract common aspects of pros and cons for semantic filtering based a personalized trade off space," describes a dynamic, natural-language-based filtering system. This is the mechanism by which the human user interacts with the outputs of Component 2.

- **Technical Architecture:** The most common and effective implementation of this is a **Retrieval-Augmented Generation (RAG)** framework. In this architecture:
    1. The "match trade-off" reports (generated by Component 2) become the *knowledge base* (the "Retrieval" part).
    2. The user's "personalized trade off space"—expressed in natural language (e.g., "I am willing to trade off industry experience for strong leadership potential")—becomes the *query* (the "Augmentation" and "Generation" part).
- **In Recommendations:** This is already in use. and describe movie recommendation

systems that "graciously handle user preferences provided... via natural language". A user can type, "I want a mind-bending sci-fi thriller like Inception". The system semantically understands this query, filters its (Component 1) profiled inventory, and ranks the results. details an LLM-powered system that integrates "semantic understanding with user preferences" to provide *cross-genre suggestions*.

- **Solving the "Filter Bubble" and Enabling Serendipity:** This component provides a powerful solution to the critical "overspecialization" failure of traditional systems. Content-based filtering can only recommend items similar to what a user has already seen. Collaborative filtering can only recommend what is popular within a user's cluster. This new "semantic filtering" allows for *serendipity*. A user can filter on *concepts* ("witty dialogue") rather than *genres* ("Comedy"), allowing the system to find novel, unexpected, yet highly relevant matches.

Case Study: Personalized Recommendations (Movies and Media)

Applying this proposed 3-stage framework to personalized movie recommendations demonstrates its significant advantages over existing methods.

- **Current State:** The dominant model in media recommendation is Collaborative Filtering (CF) , often implemented with techniques like Matrix Factorization. These systems are built on a large *user-item interaction matrix* (e.g., user ratings).
- **Trade-offs of Current State:** These methods are notoriously data-hungry and suffer from two core problems:
    1. **Cold Start Problem:** They cannot recommend items to new users (no interaction history) or recommend new items (no one has interacted with them). 2. **Data Sparsity:** The user-item matrix is, by nature, mostly empty (most users have not rated most items), which leads to weak and inaccurate recommendations.
- **Applying the 3-Stage LLM Framework:**
    - **Component 1 (Profiling):** Instead of relying on a sparse *ratings matrix*, the LLM would *ingest all of a user's unstructured reviews*. It would build a rich, semantic profile that understands the *nuance* of their preferences (e.g., "User loves complex anti-heroes and films with high-concept sci-fi, but dislikes slow-paced narratives"). This solves the **cold start problem** for *items*, as a new movie's plot summary and reviews can be profiled instantly.
    - **Component 2 (Trade-offs):** For a potential match (e.g., *Blade Runner 2049*), the LLM would generate a *contextualized explanation*. This explanation is the "trade-off" report.
        - **Example Justification:** "Based on your profile, here are the trade-offs for *Blade Runner 2049*:
            - **Pro:** You loved *Inception* and *Arrival* for their 'mind-bending' sci-fi concepts. This film shares that high-concept, philosophical DNA.
            - **Con:** Your reviews often mention you dislike 'slow-paced narratives.' This film is deliberately paced and very long, which you may find challenging."
        - **Research Validation:** The 2025 user study in *proves* this approach works. It found that "contextualized explanations (i.e., explanations that incorporate users' past behaviors)" were highly effective, "foster[ed] trust," and

"increase[d] users' intentions to watch". * **Component 3 (Filtering):** The user can now use the "personalized trade-off space" to query in natural language.
    - **Example Query:** "I'm in the mood for something with witty dialogue like *Knives Out*, but set in the 1950s."
    - **Research Validation:** This is precisely what and describe. The LLM handles the free-form text query, finds semantically relevant matches, and filters the inventory, providing a conversational and highly personalized experience.
- **Future Outlook:** This framework is the foundation for the next frontier. As research from the RecSys 2025 conference and shows, the field is moving toward *generative* and *agentic* systems. The LLM will not just *recommend* a playlist; it will *generate* a novel playlist ("Language Model-Based Playlist Generation Recommender System") and explain its choices.

Expanding the Framework: Applications in Other High-Stakes Domains

The 3-stage framework (Profile -> Trade-offs -> Filtering) is a powerful and generalizable architecture for any domain where matches are complex, nuanced, and buried in unstructured text. The research provides several powerful, real-world examples.

A. Precision Medicine: Patient-to-Clinical-Trial Matching

- **The Problem:** Patient recruitment is a "major bottleneck" in clinical trials. Matching patients is difficult because eligibility criteria are complex and patient data is split between structured records and *unstructured physician notes*.
- **The Framework in Action: "TrialMatchAI" :** This 2025 system is a *perfect implementation* of the proposed architecture.
    1. **Component 1 (Profiling):** It processes "heterogeneous clinical data," including structured records and "unstructured physician notes," to create a comprehensive patient profile.
    2. **Component 2 (Trade-offs):** It performs "criterion-level eligibility assessments" and uses "medical Chain-of-Thought reasoning" to generate *explainable outputs with traceable decision rationales*. This is the "match_trade off" report for the doctor.
    3. **Component 3 (Filtering):** The physician can then filter and review the ranked list of trials, which have been "re-ranked for criterion-level relevance".
- **Validated Impact:** This system is not theoretical. A pilot study for "TrialGPT" found it "can reduce patient screening time by 42.6%," accelerating medical research.

B. LegalTech: Semantic Matching for Case Law and Precedents

- **The Problem:** Keyword search in legal research fails. Lawyers need to find *semantically similar* concepts, not just lexically identical words. provides a critical example: a lawyer searching for "year-end bonus" would miss a key precedent where the judge used the term "annual performance bonus."
- **The Framework in Action:**
    1. **Component 1 (Profiling):** LLMs are used to read and generate AI-summaries for millions of legal opinions.

2. **Component 2 (Trade-offs):** A system like "descrybe.ai" provides an *AI-generated summary* of the case and a *match score* explaining *why* it matches the user's query. AI-powered tools *pinpoint* the "best case for a particular point of law".
3. **Component 3 (Filtering):** The user's "personalized trade off space" *is* their natural language query , which can be a complex fact pattern.

C. Human Capital: Mentor-to-Mentee Pairing

- **The Problem:** Manually matching mentors and mentees is slow, inefficient, and often sub-optimal.
- **The Framework in Action:**
    1. **Component 1 (Profiling):** The "TCH Mentor-Matching" project uses LLMs to *summarize* mentor CVs and mentee profiles.
    2. **Component 2 (Trade-offs):** highlights the system's power. It finds a "needle-in-a-haystack" match by identifying a *specific shared interest* ("vertical farming") buried deep within two different resumes, a detail a human would likely miss.
    3. **Component 3 (Filtering):** The LLM prompt itself *is* the personalized trade-off space. provides an example prompt: "Find best fit... Match manager... with higher level execs... If neither business unit nor organization have a match, next best fit by business unit.".

D. B2B/Enterprise: Semantic Partnering and Client Matching

- **The Problem:** Identifying new business partners, clients, or suppliers based on a complex, semantic understanding of their needs and capabilities.
- **The Framework in Action:**
    1. **Component 1 (Profiling):** DoorDash creates "rich, narrative-style profiles" for *merchants* (B2B partners), not just consumers.
    2. **Component 2 (Trade-offs):** and describe the complex task of matching *supplier price lists* to *internal SKU directories*. LLMs can understand attribute-level matches (e.g., distinguishing "Macallan 12-year" from "Macallan 18-year") that fuzzy matching and embeddings fail on.
    3. **Component 3 (Filtering):** shows that layering a semantic "Knowledge Graph" over a SQL database *triples* the accuracy of LLM-based query-answering for complex business questions (from 16.7% to 54.2%), validating the power of semantic filtering.

Critical Challenges and Implementation Barriers

The viability of the proposed framework is not a given. A naive implementation, as specified, would fail. The framework's success is contingent on overcoming three enterprise-ending challenges: scalability, bias, and privacy.

A. The Scalability Bottleneck: The O(n^2) Problem of Pairwise Comparisons

- **The Proposal:** "consider match(in, jn) for each I and each j."
- **The Problem:** This "exhaustive" or "brute-force" comparison is computationally and financially non-viable. This is a well-known *quadratic complexity* problem, or O(n^2).
    - **The Math:** For a system with N items, the number of comparisons scales quadratically. If a company has 1,000 candidates (i) for 1,000 open jobs (j), the system must perform 1,000 * 1,000 = **1,000,000** pairwise LLM comparisons.
    - **The Consequence:** This is identified in 2024-2025 research as a "substantial bottleneck," "intractable," and a source of "poor scalability" due to its "quadratic query complexity".
    - **The Cost:** Each of those 1,000,000 comparisons is an LLM API call. The financial cost would be astronomical.
- **The Implication:** This is the single greatest *technical* flaw in the proposed architecture. The "exhaustive" comparison, while ideal in theory, is impossible in practice. This *forces* a different, more intelligent architecture (see Section VII.A). While alternatives like "Knockout Assessment" or pointwise ranking are being researched, the O(n^2) cost of full pairwise comparison remains a prohibitive barrier.

B. The "Illusion of Thinking": The New Face of Algorithmic Bias

- **The Assumption:** The query implies that LLMs, being more advanced, will be *less* biased than the old systems.
- **The Reality:** This assumption is dangerously false. The research is clear: general-purpose LLMs *do not solve bias; they obfuscate it.*
    - **Reasoning Failures:** LLM reasoning can be an "illusion". They are notoriously prone to *positional bias* (e.g., a "judge" LLM favoring the first option in a pair, regardless of content) and *order inconsistency* , making their "trade-off" judgments unreliable.
    - **Severe Racial and Gender Bias:** The evidence on LLMs in hiring is damning.
        - Amazon's early AI was famously biased against women.
        - A 2024 University of Washington study screening 550 resumes with three state-of-the-art LLMs found they **favored white-associated names 85% of the time** and **NEVER favored Black male-associated names over white male names.**
        - A 2025 PNAS study confirmed this, finding LLMs award *lower* assessment scores to Black male candidates, resulting in a **1.4 percentage-point lower hiring probability** for *otherwise identical candidates*.
    - **Implicit & Intersectional Bias:** The bias is multi-layered. Even when models are tuned to reduce *explicit* race/gender bias , they retain *implicit biases*. notes a "preference for elite education." The bias is also *intersectional* : the models penalize "Black male" names differently and more severely than "Black female" names.
- **The Critical Finding: "Domain-Specific" vs. "General-Purpose":** This is the most important finding in this report. and present a direct comparison from the AI-hiring company Eightfold.ai. They benchmarked their *proprietary, domain-specific, supervised* "Match Score" model against *general-purpose foundational LLMs* (OpenAI, Google, Anthropic) on 10,000 real-world candidate-job pairs.
    - **The Result:** The domain-specific model was *more accurate* (ROC AUC 0.85 vs. 0.77 for the best LLM) and *significantly more fair* (minimum race-wise impact ratio of **0.957** [near-parity] vs. **0.809** or lower for the LLMs).

    - **The Implication:** This *proves* that simply "applying LLMs" (the naive proposal) is the *wrong* approach. It results in a system that is *less accurate and more biased*. The *correct* approach is to use a *bespoke, supervised, domain-specific model with "extensive fairness safeguards" built in*. Bias can stem from names or from resume content itself.

Table 2: Risk-Benefit Analysis: General-Purpose LLMs vs. Domain-Specific Models in Hiring

Model Type	General-Purpose LLMs (e.g., OpenAI, Google, Anthropic)	Proprietary, Domain-Specific Supervised Model (e.g., "Match Score")
Accuracy (ROC AUC)	0.77 (or lower)	**0.85**
Fairness (Min. Race-Impact Ratio)	0.809 (or lower) - *Highly Biased*	**0.957** - *Near Parity*
Fairness (Intersectional Impact)	0.773 (or lower)	**0.906**
Key Takeaway	Off-the-shelf LLMs are *less accurate* and *significantly more biased* than the systems they are meant to replace.	A bespoke model with "safeguards built in" can achieve both state-of-the-art accuracy *and* fairness.

C. The Privacy Mandate: Processing High-Stakes Sensitive Data

- **The Problem:** The proposed framework requires "consuming" the most sensitive data imaginable:
    - **Dating:** All profile information, private DMs, and swipe behavior.
    - **Hiring:** All candidate resumes, cover letters, and recruiter notes.
    - **Medicine:** Patient Electronic Health Records (EHRs), including unstructured physician notes.
- **The Risk:** This creates a massive attack surface and a legal/compliance nightmare.
    - **Legal Risk:** The European Data Protection Board (EDPB) explicitly identifies "processing sensitive data" for a "sensitive & impactful purpose" (like automated *hiring decisions*) as a "High level Risk" factor that can "negatively impact individuals".
    - **Technical Risk:** LLMs are known to *memorize and leak* sensitive personal data from their training sets. They are also vulnerable to *prompt injection* attacks, where a malicious user could craft a resume to exfiltrate data from the system.
- **The Architectural Paradox:** This creates a dilemma. The *most powerful* LLMs are closed-source, third-party APIs. However, no hospital (HIPAA), bank, or HR department (GDPR) can *ethically or legally* send all their unanonymized, sensitive patient/candidate data to a third-party API for processing.
- **The Solution:** The only viable architecture is one built for *privacy*. This means using *locally deployable* models or *private-cloud* instances. The "TrialMatchAI" system provides the blueprint, explicitly noting it is designed for "secure local deployment" to ensure patient data privacy and compliance. The technology is "too personally integrated" to be controlled by a third party.

Strategic Recommendations and Future Outlook

The proposed framework is conceptually sound but naively specified. The following strategic recommendations are required to transform it from a high-risk theoretical concept into a viable, defensible, and effective system.

A. Recommendation 1: Mitigate Quadratic Complexity with a Hybrid Architecture

- **The Problem:** The "exhaustive" O(n^2) pairwise comparison (Component 2) is computationally intractable.
- **The Solution:** Do not build an exhaustive system. Implement a *hybrid, multi-stage "retrieve-and-rerank" architecture*.
    1. **Stage 1: Retrieval (Scalable & Cheap):** Use a *scalable*, low-cost algorithm to "retrieve" the Top-K (e.g., Top 100) most promising matches from the entire database. This could be a traditional model (e.g., SLIM) or, more likely, a *semantic vector search*. This stage turns the O(N) problem into a manageable O(k).
    2. **Stage 2: Reranking & Generation (Intensive & Expensive):** Apply the *expensive* "Component 2" (the "match_trade offs" LLM generation) *only* to this small set of Top-K candidates. This makes the computationally "intractable" problem suddenly tractable and cost-effective.

B. Recommendation 2: A Framework for Auditing and Mitigating LLM-Native Bias

- **The Problem:** General-purpose, "off-the-shelf" LLMs are *not* neutral. They are *less accurate* and *more biased* than specialized models for high-stakes domains like hiring.
- **The Solution:** Do not use a general-purpose LLM.
    1. **Invest in Domain-Specific Fine-Tuning:** The solution is to create a *bespoke, supervised, domain-specific model*. This model must be *fine-tuned* on domain-specific data.
    2. **Audit for Accuracy and Fairness:** This model *must* be rigorously audited against a ground-truth dataset, benchmarking it for both *predictive accuracy* (e.g., ROC AUC) and *fairness* (e.g., "impact ratio").
    3. **Audit Intersectional Bias:** The audit must be *intersectional*, checking for bias not just on "race" or "gender" but on "Black male" vs. "white female" vs. "Black female" and *implicit* biases like "elite education". This is the *only* ethically and legally defensible path forward.

C. Recommendation 3: Architect for "Privacy-by-Design"

- **The Problem:** The system's inputs (resumes, chats, medical records) are highly sensitive and regulated. Using third-party, closed-source APIs is not a viable option.
- **The Solution:** The system architecture *must* be "privacy-by-design."
    1. **Use Local or Private Cloud Deployment:** The architecture should be built using *open-source, locally deployable models* or deployed in a *secure, private cloud*.
    2. **Emulate "TrialMatchAI":** The "TrialMatchAI" system provides the blueprint, explicitly stating it is "designed for... secure local deployment" to ensure patient data privacy and compliance. This must be a core, non-negotiable feature of the

framework.

D. Concluding Analysis: The Viability of the LLM-Powered "Trade-Off" Model

The core insight that the era of the simple "match score" is over is correct. The proposed 3-stage framework (Semantic Profiling -> Trade-off Analysis -> Personalized Filtering) is a visionary and sound architecture, validated by real-world applications in medicine, law, and enterprise.

However, the *implementation* is fraught with critical, enterprise-ending challenges. A naive approach—applying an off-the-shelf public LLM to an exhaustive pairwise comparison—is a technical, financial, and ethical "time bomb." It will be computationally intractable, prohibitively expensive, and will expose the organization to massive legal liability from its deeply-biased and non-private operations.

The vision is only viable if it pivots:

1. From **Exhaustive** to **Hybrid**.
2. From **General-Purpose** to **Domain-Specific & Audited**.
3. From **Public API** to **Private & Secure**.

By addressing these three challenges, the proposed framework moves from a "black box" *gatekeeper* to an *explainable, co-pilot* system. This is the true, high-value promise of LLMs: not to replace human judgment, but to augment it with synthesized, semantic, and transparent intelligence.

Works cited

1. TrialMatchAI: An End-to-End AI-powered Clinical Trial ... - arXiv, https://arxiv.org/abs/2505.08508 2. Free Online Caselaw Searching: Descrybe.ai - Jenkins Law Library, https://www.jenkinslaw.org/blog/2024/08/29/free-online-caselaw-searching-descrybeai 3. Estimating the Error of Large Language Models at Pairwise Text Comparison - arXiv, https://arxiv.org/html/2510.22219v1 4. Quadratic Complexity in LLMs: Why AI Struggles with Long Texts ..., https://nat.io/blog/quadratic-complexity-llms 5. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators - arXiv, https://arxiv.org/html/2403.16950v5 6. AI tools show biases in ranking job applicants' names according to ..., https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/ 7. Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions - arXiv, https://arxiv.org/html/2507.02087v1 8. Measuring gender and racial biases in large language models ..., https://academic.oup.com/pnasnexus/advance-article/doi/10.1093/pnasnexus/pgaf089/8071848 9. Data-Hungry Dating Apps Are Worse Than Ever for Your Privacy - Mozilla Foundation, https://www.mozillafoundation.org/en/privacynotincluded/articles/data-hungry-dating-apps-are-worse-than-ever-for-your-privacy/ 10. AI Privacy Risks & Mitigations – Large Language Models (LLMs), https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf 11. State of Recommender Systems in 2025: Algorithms, Libraries, and Trends - Reddit, https://www.reddit.com/r/recommendersystems/comments/1iwwxpr/state_of_recommender_systems_in_2025_algorithms/ 12. LLM Optimization Unlocks Real-Time Pairwise Reranking - arXiv,

https://arxiv.org/html/2511.07555v1 13. I'm convinced now that “personal LLMs” are going to be a huge thing - Reddit, https://www.reddit.com/r/LocalLLaMA/comments/16au3ga/im_convinced_now_that_personal_llms_are_going_to/ 14. The best dating apps of 2025 to cure 'app fatigue' | Mashable, https://mashable.com/roundup/best-dating-apps-2025 15. What is the latest information out there on how the algorithm's work on Tinder, Bumble and Hinge and how can a man take advantage of the algorithm to gain lot of matches? : r/SwipeHelper - Reddit, https://www.reddit.com/r/SwipeHelper/comments/1htl3xi/what_is_the_latest_information_out_there_on_how/ 16. Unravelling The Secrets Of The Latest Tinder Algorithm 2025 | Appscrip Blog, https://appscrip.com/blog/secrets-of-the-latest-tinder-algorithm-2024/ 17. Top Dating App Development Trends 2025 - Fulminous Software, https://fulminoussoftware.com/top-dating-app-development-trends-2025 18. Best Online Dating Apps And Sites In 2025 - Forbes, https://www.forbes.com/health/dating/best-online-dating-websites/ 19. (PDF) DATING THROUGH THE FILTERS - ResearchGate, https://www.researchgate.net/publication/351311593_DATING_THROUGH_THE_FILTERS 20. Cupid trades arrows for algorithms: dissecting our modern dating scene - Rostra Economica, https://www.rostraeconomica.nl/post/cupid-trades-arrows-for-algorithms-dissecting-our-modern-dating-scene 21. Dating apps' darkest secret: their algorithm - IE HST Rewire Magazine, https://rewire.ie.edu/dating-apps-darkest-secret-algorithm/ 22. Hinge and Its Implementation of the Gale–Shapley algorithm | Hacker News, https://news.ycombinator.com/item?id=31748967 23. Dating Through the Filters | Montreal AI Ethics Institute, https://montrealethics.ai/dating-through-the-filters/ 24. DATING THROUGH THE FILTERS | Social Philosophy and Policy ..., https://www.cambridge.org/core/journals/social-philosophy-and-policy/article/dating-through-the-filters/EA64BE27CD7D2A1749D712A5E179828D 25. Navigating an Applicant Tracking System (ATS) - NDSU Career and Advising Center, https://career-advising.ndsu.edu/navigating-an-applicant-tracking-system-ats/ 26. Applicant Tracking Systems: Everything You Need to Know - Jobscan, https://www.jobscan.co/applicant-tracking-systems 27. ATS Resume Optimization: The Ultimate 2025 Guide to Getting Past the Bots, https://blog.theinterviewguys.com/ats-resume-optimization/ 28. Full Guide to Optimizing Resume Keywords to Pass ATS Screening : r/jobsearchhacks, https://www.reddit.com/r/jobsearchhacks/comments/1j530wc/full_guide_to_optimizing_resume_keywords_to_pass/ 29. What is an Applicant Tracking System? | Workday US, https://www.workday.com/en-us/topics/hr/applicant-tracking-system.html 30. What is an Applicant Tracking System (ATS)? A Full 2025 Guide - Oleeo, https://www.oleeo.com/blog/what-is-an-applicant-tracking-system-ats/ 31. A Study of Reciprocal Job Recommendation for College Graduates Integrating Semantic Keyword Matching and Social Networking - MDPI, https://www.mdpi.com/2076-3417/13/22/12305 32. Beyond Keywords: AI Semantic Search & Headhunting - shortlistd.io, https://www.shortlistd.io/blog/beyond-keywords-how-semantic-search-enables-the-headhunting-revolution 33. Beyond Keywords: How to Make AI Recruitment Tools Useful for Your Business - Addepto, https://addepto.com/blog/beyond-keywords-how-to-make-ai-recruitment-tools-actually-work-for-your-company/ 34. Upturn | Help Wanted: An Examination of Hiring Algorithms, Equity ..., https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf

35. Applicant Tracking System Statistics (Updated for 2025) - SSR - SelectSoftware Reviews, https://www.selectsoftwarereviews.com/blog/applicant-tracking-system-statistics 36. Mastering Recruitment in 2025: How to Solve the Biggest Hiring Roadblocks, https://www.recruitmentsmart.com/blogs/mastering-recruitment-in-2025-how-to-solve-the-biggest-hiring-roadblocks 37. What Is Semantic Search in Recruiting (and Why It Matters) - Stardex, https://www.stardex.com/blog/what-is-semantic-search 38. Resume2Vec: Transforming Applicant Tracking Systems with Intelligent Resume Embeddings for Precise Candidate Matching - MDPI, https://www.mdpi.com/2079-9292/14/4/794 39. The Great Hiring Reset: Why 2025 Will Break Traditional Recruitment - Recrew AI, https://www.recrew.ai/blog/why-2025-will-break-traditional-recruitment 40. AI-Driven Candidate Screening: The 2025 In-Depth Guide, https://www.herohunt.ai/blog/ai-driven-candidate-screening-the-2025-in-depth-guide 41. How Large Language Models Interpret Job Descriptions - Resumly.ai, https://www.resumly.ai/blog/how-large-language-models-interpret-job-descriptions 42. ATS Resume Optimization with AI: Guide to Pass the Filters - JobWinner.ai, https://jobwinner.ai/blog/how-to-pass-ats-filters-with-ai/ 43. How AI Is Changing Hiring in 2025: What Job Seekers Need to Know | Sensei AI, https://www.senseicopilot.com/blog/how-ai-is-changing-hiring-in-2025 44. Large Language Models: Evolution, State of the Art in 2025, and Business Impact | Proffiz, https://proffiz.com/large-language-models-in-2025/ 45. Improving Recommendation Systems & Search in the Age of LLMs - Eugene Yan, https://eugeneyan.com/writing/recsys-llm/ 46. A Comparison of All Leading LLMs - AI-PRO.org, https://ai-pro.org/learn-ai/articles/a-comprehensive-comparison-of-all-llms 47. User Profile with Large Language Models: Construction, Updating, and Benchmarking, https://arxiv.org/html/2502.10660v1 48. Tailoring LLM Responses to Individual User Preferences - Sapien, https://www.sapien.io/blog/tailoring-llm-responses-to-individual-user-preferences-and-needs 49. LLM-Powered Parsing and Analysis of Semi-Structured & Structured Documents, https://towardsdatascience.com/llm-powered-parsing-and-analysis-of-semi-structured-structured-documents-f03ac92f063e/ 50. Finding Matches: A Guide to List Matching with LLM | by Gregory Zem | Medium, https://medium.com/@mne/finding-matches-a-guide-to-list-matching-with-llm-2ae54fd0985e 51. Content-based filtering advantages & disadvantages | Machine Learning, https://developers.google.com/machine-learning/recommendation/content-based/summary 52. What are the limitations of content-based filtering? - Milvus, https://milvus.io/ai-quick-reference/what-are-the-limitations-of-contentbased-filtering 53. Comparison Between Collaborative Filtering and Content-Based Filtering - ResearchGate, https://www.researchgate.net/publication/365494125_Comparison_Between_Collaborative_Filtering_and_Content-Based_Filtering 54. LLM-based User Profile Management for Recommender ... - arXiv, https://arxiv.org/abs/2502.14541 55. [2508.08454] Temporal User Profiling with LLMs: Balancing Short-Term and Long-Term Preferences for Recommendations - arXiv, https://arxiv.org/abs/2508.08454 56. From Millions of Tweets to Actionable Insights: Leveraging LLMs for User Profiling - arXiv, https://arxiv.org/html/2505.06184v1 57. Profile Generation with LLMs: Understanding consumers, merchants, and items - DoorDash, https://careersatdoordash.com/blog/doordash-profile-generation-llms-understanding-consumers-merchants-and-items/ 58. arxiv.org, https://arxiv.org/html/2504.02870v1 59. AI Recruitment 2025: The Extremely In-Depth Expert Guide (10k words) - HeroHunt.ai, https://www.herohunt.ai/blog/ai-recruitment-2025-the-extremely-in-depth-expert-guide-10k-word

s 60. Ultimate Guide to AI Candidate Matching 2025 - Skillfuel, https://www.skillfuel.com/ultimate-guide-to-ai-candidate-matching/ 61. How to use LLMs in recruitment: a practical guide - HeroHunt.ai, https://www.herohunt.ai/blog/how-to-use-llms-in-recruitment 62. 7 Recruiting Tasks AI Will Revolutionize in 2025 - SmartRecruiters, https://www.smartrecruiters.com/resources/article/recruiting-tasks-ai-will-revolutionize/ 63. AI in Recruitment: Pros and Cons in 2025 - AMS - Alexander Mann Solutions, https://www.weareams.com/blog/ai-in-recruitment/ 64. Seven limitations of Large Language Models (LLMs) in recruitment technology - Textkernel, https://www.textkernel.com/learn-support/blog/seven-limitations-of-llms-in-hr-tech/ 65. AI Recruiting in 2025: The Complete Guide - Connect2BPO, https://connect2bpo.com/ai-recruiting-in-2025-the-complete-guide/ 66. On explaining recommendations with Large Language Models: a review - Frontiers, https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1505284/full 67. On explaining recommendations with Large Language Models: a review - PMC, https://pmc.ncbi.nlm.nih.gov/articles/PMC11808143/ 68. Can LLM-Generated Textual Explanations Enhance Model Classification Performance? An Empirical Study - arXiv, https://arxiv.org/html/2508.09776v1 69. Chapter 4 Large Language Model Driven Recommendation - arXiv, https://arxiv.org/html/2408.10946v1 70. REGEN: Empowering personalized recommendations with natural language, https://research.google/blog/regen-empowering-personalized-recommendations-with-natural-language/ 71. Contextualizing Recommendation Explanations with LLMs: A User ..., https://arxiv.org/abs/2501.12152 72. How AI semantic search with LLMs is redefining enterprise search - Pretius, https://pretius.com/blog/ai-semantic-search-with-llm 73. Beyond RAG: Precision Filtering in a Semantic World - Towards Data Science, https://towardsdatascience.com/beyond-rag-precision-filtering-in-a-semantic-world-333d332c2d45/ 74. Semantic Decomposition and Selective Context Filtering - arXiv, https://arxiv.org/html/2502.14048v1 75. A Language-Driven Framework for Improving Personalized Recommendations: Merging LLMs with Traditional Algorithms - arXiv, https://arxiv.org/html/2507.07251v1 76. Movie Recommendation System using LLM model - Ready Tensor, https://app.readytensor.ai/publications/movie-recommendation-system-using-llm-model-1KxdQ1nJ1DQ1 77. Personalized Recommendation Systems Powered By Large Language Models Integrating Semantic Understanding and User Preferences, https://ijirem.org/DOC/6-Personalized-Recommendation-Systems-Powered-By-Large-Language-Models-Integrating-Semantic-Understanding-and-User-Preferences.pdf 78. What is content-based filtering? - IBM, https://www.ibm.com/think/topics/content-based-filtering 79. Collaborative filtering | Machine Learning - Google for Developers, https://developers.google.com/machine-learning/recommendation/collaborative/basics 80. Building Recommendation Systems in 2025 | Boost User Engagement - Rapid Innovation, https://www.rapidinnovation.io/post/how-to-build-a-recommendation-system-process-features-costs 81. Why Are Recommendation Systems Hard to Implement? Insights for Dating and E-commerce : r/learnmachinelearning - Reddit, https://www.reddit.com/r/learnmachinelearning/comments/1h9h5fc/why_are_recommendation_systems_hard_to_implement/ 82. Limitations Of Collaborative Filtering - Meegle, https://www.meegle.com/en_us/topics/recommendation-algorithms/limitations-of-collaborative-filtering 83. An evaluation of LLMs for generating movie reviews: GPT-4o, Gemini-2.0 and DeepSeek-V3,

https://www.researchgate.net/publication/392334164_An_evaluation_of_LLMs_for_generating_movie_reviews_GPT-4o_Gemini-20_and_DeepSeek-V3 84. RecSys 2025 - Accepted Contributions, https://recsys.acm.org/recsys25/accepted-contributions/ 85. Second Workshop on Generative AI for Recommender Systems and Personalization (2025), https://genai-personalization.github.io/GenAIRecP2025 86. RecSys 2025 - Tutorials - ACM, https://recsys.acm.org/recsys25/tutorials/ 87. From Traditional Recommender Systems to Generative AI: Redefining Personalized Recommendations | by Sia AI - Medium, https://medium.com/sia-ai/from-traditional-recommender-systems-to-generative-ai-redefining-personalized-recommendations-d9f892460f46 88. Enhancing Patient-Trial Matching With Large Language Models: A Scoping Review of Emerging Applications and Approaches - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC12169815/ 89. [2503.15374] Real-world validation of a multimodal LLM-powered pipeline for High-Accuracy Clinical Trial Patient Matching leveraging EHR data - arXiv, https://arxiv.org/abs/2503.15374 90. Large Language Models Help Match Patients to Clinical Trials - National Cancer Institute, https://www.cancer.gov/about-nci/organization/cbiit/news-events/news/2024/large-language-models-help-match-patients-clinical-trials 91. Designing Human-AI System for Legal Research: A Case Study of Precedent Search in Chinese Law - arXiv, https://arxiv.org/html/2504.08235v1 92. AI-Driven Legal Research and Tools - Bloomberg Law, https://pro.bloomberglaw.com/products/ai-and-bloomberg-law/ 93. AI Legal Precedent Matching Guide 2025 - Rapid Innovation, https://www.rapidinnovation.io/post/ai-agent-precedent-matching-assistant 94. Application of LLMs to Pairing Use Cases: Mentor and Mentee Matching - Medium, https://medium.com/@manishasolipuram/application-of-llms-to-pairing-use-cases-mentor-and-mentee-matching-5022c7f67552 95. Matching Mentees and Mentors - excel - Stack Overflow, https://stackoverflow.com/questions/72806809/matching-mentees-and-mentors 96. LiuzLab/tch-mentormatching - GitHub, https://github.com/LiuzLab/tch-mentormatching 97. Find the Perfect Mentor for Each Mentee Using AI - ASAE, https://www.asaecenter.org/resources/articles/an_plus/2024/10-october/find-the-perfect-mentor-for-each-mentee-using-ai 98. Top AI Platforms for Semantic B2B Search - Landbase, https://www.landbase.com/blog/top-ai-platforms-for-semantic-b2b-search 99. LLMs and semantic models: Complementary technologies for enhanced Business Intelligence - Tabular Editor 3, https://tabulareditor.com/blog/llms-and-semantic-models-complementary-technologies-for-enhanced-business-intelligence 100. Semantic Layer as the Data Interface for LLMs - dbt Labs, https://www.getdbt.com/blog/semantic-layer-as-the-data-interface-for-llms 101. [2405.05894] Efficient LLM Comparative Assessment: a Product of Experts Framework for Pairwise Comparisons - arXiv, https://arxiv.org/abs/2405.05894 102. Top Platforms For Side-By-Side LLM Comparison | Prompts.ai, https://www.prompts.ai/en/blog/platforms-side-by-side-llm-comparison 103. The Hidden Cost of LLM-as-a-Judge: When More Evaluation Means Less Value, https://www.soumendrak.com/blog/llm-evals/ 104. LLM Cost Optimization: Complete Guide to Reducing AI Expenses by 80% in 2025, https://ai.koombea.com/blog/llm-cost-optimization 105. [2410.06550] Investigating Cost-Efficiency of LLM-Generated Training Data for Conversational Semantic Frame Analysis - arXiv, https://arxiv.org/abs/2410.06550 106. How to Build Cost-Effective Semantic Search with LLMs - TiDB, https://www.pingcap.com/article/cost-effective-semantic-search-llms/ 107. 15 Proven Strategies to Reduce LLM Costs Without Sacrificing Performance - ARON HACK, https://aronhack.com/15-proven-strategies-to-reduce-llm-costs-without-sacrificing-performance/ 108. Knockout LLM Assessment: Using Large Language Models for Evaluations through

Iterative Pairwise Comparisons - ACL Anthology, https://aclanthology.org/2025.gem-1.10.pdf 109. J1: Incentivizing Thinking in LLM-as-a-Judge via Reinforcement Learning - arXiv, https://arxiv.org/html/2505.10320v2 110. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity - Apple Machine Learning Research, https://machinelearning.apple.com/research/illusion-of-thinking 111. Diagnosing Bias and Instability in LLM Evaluation: A Scalable Pairwise Meta-Evaluator, https://www.mdpi.com/2078-2489/16/8/652 112. What Is Algorithmic Bias? - IBM, https://www.ibm.com/think/topics/algorithmic-bias 113. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC11937954/ 114. Invisible Filters: Cultural Bias in Hiring Evaluations Using Large Language Models - arXiv, https://arxiv.org/html/2508.16673v1 115. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race ..., https://aclanthology.org/2025.naacl-industry.55/ 116. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education - arXiv, https://arxiv.org/abs/2503.19182 117. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education - ACL Anthology, https://aclanthology.org/2025.naacl-industry.55.pdf 118. Two Tickets are Better than One: Fair and Accurate Hiring Under Strategic LLM Manipulations - arXiv, https://arxiv.org/pdf/2502.13221? 119. No Thoughts Just AI: Biased LLM Recommendations Limit Human Agency in Resume Screening - arXiv, https://arxiv.org/html/2509.04404v1 120. Dangers of using LLMs to rank and even select candidates | ep107 - YouTube, https://www.youtube.com/watch?v=R8z7FsSMyyA 121. Instructions for *ACL Proceedings - arXiv, https://arxiv.org/html/2406.12232v2 122. No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy - arXiv, https://arxiv.org/html/2509.04404v2 123. Robustly Improving LLM Fairness in Realistic Settings via Interpretability - arXiv, https://arxiv.org/html/2506.10922v1 124. 1 Introduction - arXiv, https://arxiv.org/html/2509.00462v2 125. Annual Meeting of the Association for Computational Linguistics (2025) - ACL Anthology, https://aclanthology.org/events/acl-2025/ 126. Large Language Models Are Biased Because They Are Large Language Models | Computational Linguistics - Open Journal Systems, https://submissions.cljournal.org/index.php/cljournal/article/view/2990 127. Gender, race, and intersectional bias in AI resume screening via language model retrieval, https://www.brookings.edu/articles/gender-race-and-intersectional-bias-in-ai-resume-screening-via-language-model-retrieval/ 128. Identifying and Mitigating Privacy Risks Stemming from Language Models - arXiv, https://arxiv.org/html/2310.01424v2 129. How Indeed builds and deploys fine-tuned LLMs on Amazon SageMaker, https://aws.amazon.com/blogs/machine-learning/how-indeed-builds-and-deploys-fine-tuned-llms-on-amazon-sagemaker/ 130. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learn - The University of Texas at Dallas, https://www.utdallas.edu/~sriraam.natarajan/Papers/HRSPreprint.pdf 131. Job Recommendations: Benchmarking of Collaborative Filtering Methods for Classifieds, https://www.mdpi.com/2079-9292/13/15/3049 132. The 10 Best Semantic Search APIs in 2025 | Shaped Blog, https://www.shaped.ai/blog/the-10-best-semantic-search-apis-in-2025 133. Purely Semantic Indexing for LLM-based Generative Recommendation and Retrieval - arXiv, https://arxiv.org/html/2509.16446v1 134. Reducing LLM Costs and Latency via Semantic Embedding Caching - arXiv, https://arxiv.org/html/2411.05276v2 135. A three-step design pattern for specializing LLMs | Google Cloud Blog, https://cloud.google.com/blog/products/ai-machine-learning/three-step-design-pattern-for-specializing-llms 136. Augmented Fine-Tuned LLMs for Enhanced Recruitment Automation - arXiv,

https://arxiv.org/html/2509.06196v1