

Here's a clear, build-ready specification for **AI Mates** in English—covering all **13 points**: data model, pipeline, prompts, scoring, QA, privacy/safety, architecture, example I/O, UX, edge cases, pseudocode, test plan, and the core AI capabilities.

Vision

Before matching, a user sees **(2)** a *warm, generous* short summary of a simulated conversation and **(3)** a 1–10 compatibility score. The **(1)** simulated conversation is generated internally to power the summary (and potential future features) but is **not** shown pre-match.

1) Data Model (Profiles & Preferences)

Profile (User)

- user_id (UUID, pseudonymized)
- age (Integer), pronouns (free text), location_city/region, distance_radius_km
- **Free-text fields**: bio, interests, hobbies, values, fun_facts, conversation_starters
- **Structured fields**: education, occupation, languages (with CEFR), lifestyle (smoking, alcohol, sport, diet), religion (optional), pets, wants_kids (optional)
- **Match preferences**: preferred_age_range, max_distance_km, dealbreakers (list), hard_no_topics (list), looking_for (relationship type), language_pref_order (ranking)
- **Safety flags** (from moderation): is_minor (must be false), nsfw_flag, spam_risk_score, policy_violations
- **Consent**: llm_processing_consent (bool, required), data_usage (A/B/Full), show_summary_pre_match (bool)

Derived features

- Text embeddings per field (e.g., bio_embed, values_embed, interests_embed)
 - Topic tags (LLM/classifier-assisted)
 - **Evidence pointers**: references to specific profile snippets permitted in summaries/rationales.
-

2) Matching Pipeline (End-to-End)

Input: Profile A, Profile B

Output: (1) simulated conversation (internal), (2) generous short summary (visible), (3) compatibility score 1–10 (visible)

1. Eligibility & Filters

- Both ≥ 18 ; basic preferences met (age, distance, relationship intent, language overlap).
- Enforce dealbreakers (e.g., “no smoking” vs. “smokes regularly”).

2. Feature Fusion

- Cosine similarities on embeddings (bio, values, interests).
- Distance penalty (Haversine) vs. max_distance_km.
- Language compatibility (shared language + preference weights).
- Lifestyle compatibility (rule-based + optional embedding proximity).

3. LLM Conversation (internal)

- System prompt enforces respectful, evidence-based chat; no facts beyond profiles; 12–18 total messages (6–9 each); realistic length; concrete hooks from profiles.
- Persona brackets: each line labeled [A] or [B], in each profile’s voice.
- Safety rules: no NSFW, no inference of sensitive attributes, no PII.

4. Generous Short Summary (visible)

- 90–140 words; warm, specific, concise.
- Structure: 2–3 shared themes, 1 actionable conversation starter, 1 friendly “if you match” suggestion.
- **Evidence anchors:** light bracketed references to profile snippets (no private data).

5. Compatibility Score (visible)

- Raw score = weighted sum:
 - S_interests (0.25), S_values (0.25), S_lifestyle (0.15),
 - S_language (0.10), S_distance (0.15), S_dealbreakers (0.10; zero if violated).
- **Calibration:** isotonic regression or Platt scaling vs. ground truth (e.g., mutual likes / 5+ message chats).
- Map to integer 1–10 + confidence band (low/medium/high).

6. Moderation/Policy

- Run moderation on the summary before display (toxicity, sexual content, sensitive inference).
 - Block/replace on flags; log for review.
-

3) Prompts (English, drop-in ready)

System Prompt — Conversation Simulation

You simulate a short, authentic chat between two real people using ONLY their profiles.

Rules: friendly, respectful, no spammy small talk, no assumptions beyond profiles, no sensitive inferences.

Alternate messages:

[A] ...

[B] ...

Max 16 messages (8 per person). Each turn 1–3 sentences, grounded in explicit profile details.

Avoid clichés, sexual content, diagnosis, or lecturing.

End without a conclusion.

User Prompt — Conversation (inject inputs)

Profile A:

- Bio: {A.bio}
- Interests: {A.interests}
- Values: {A.values}
- Lifestyle: {A.lifestyle}
- Languages: {A.languages}
- Conversation starters: {A.conversation_starters}

Profile B:

- Bio: {B.bio}
- Interests: {B.interests}
- Values: {B.values}

- Lifestyle: {B.lifestyle}
- Languages: {B.languages}
- Conversation starters: {B.conversation_starters}

Begin the conversation with [A].

System Prompt — Generous Summary

Write a warm, precise short summary (90–140 words) of the conversation above.

Structure:

- Where they complement each other (2–3 points),
- One concrete conversation prompt,
- A positive, non-pushy close.

Optionally reference profile evidence in brackets (e.g., from A: "trail running").

No sensitive inferences; no advice about risky behavior.

Scoring Prompt (optional LLM-assisted instead of rule-only)

Rate the compatibility of Profile A and B from 1–10 across interests, values, lifestyle, language, distance, and dealbreakers.

Return JSON only:

```
{"score": <1-10>, "rationale": "<max 40 words>", "signals": ["...", "..."]}
```

4) Example I/O

Input (abridged)

```
{
  "A": {
    "bio": "Product designer, loves trail running & espresso.",
    "interests": ["Trail running", "UX", "Alps"],
    "values": ["Honesty", "Curiosity"],
    "languages": ["de-C1", "en-B2"],
    "lifestyle": {"smoking": "no", "alcohol": "social", "diet": "vegetarian"},
    "dealbreakers": ["Smoking"]
  }
}
```

```

},
"B": {
  "bio": "Teacher, mountain hikes, photography.",
  "interests": ["Hiking","Photography","Coffee"],
  "values": ["Empathy","Honesty"],
  "languages": ["de-B2","fr-B1"],
  "lifestyle": {"smoking":"no","alcohol":"rare","diet":"omnivore"},
  "dealbreakers": []
},
"distance_km": 12
}

```

Output (visible)

```

{
  "summary": "You both light up around mountains and good coffee: A runs trails, B plans weekend hikes — a natural fit (from A: “trail running”, from B: “mountain hikes”). Your values overlap in honesty and a curious outlook. A photo-story walk could be an easy start: B brings the camera, A suggests a route with views and an espresso stop. German works for both, and 12 km is close enough for spontaneous plans. It feels like a relaxed get-to-know-you with no pressure — maybe a short hike ending at a café?",
  "compatibility_score": 8,
  "confidence": "medium"
}

```

(The simulated conversation remains internal.)

5) Scoring Details (Rules + Embeddings)

Embeddings

- Compute per-field sentence/paragraph embeddings (bio, interests, values); aggregate per field.
- $S_{\text{interests}} = \text{cosine}(\text{mean}(A.\text{interests_embed}), \text{mean}(B.\text{interests_embed}))$
- S_{values} similar; $S_{\text{lifestyle}}$ = hybrid metric (rules + optional text embedding of lifestyle descriptions).

- S_language: 1 if shared language at $\geq B2$; 0.5 for B1; else 0.
- S_distance: 1 within $\leq \frac{1}{3}$ max_distance_km, linearly decreasing to 0 beyond max_distance_km.
- S_dealbreakers: 0 if any conflict, else 1.

Calibration

- $\text{Raw} \in [0,1] \rightarrow \text{score}_{1_10} = \text{round}(1 + 9 * \text{calib}(\text{raw}))$
 - calib via isotonic regression fitted on historical acceptance/chat outcomes.
-

6) Quality Assurance & Metrics

Offline

- ~200–500 curated profile pairs with 3+ human raters.
- Spearman correlation between raw score and human rating.
- Toxic/NSFW/inference rate of LLM outputs < 0.5%.
- Summary readability in target band; length 90–140 words.

Online (A/B)

- Pre-match preview open rate.
 - Mutual-interest rate after showing (2) & (3).
 - Chat-start rate and 24-hour reply rate.
 - Report/“uncomfortable” click rate (should decrease).
 - Calibration: expected vs. observed chat probability per score bucket.
-

7) Safety, Fairness, Legal (CH/EU Focus)

- **Consent (GDPR/DSG Art. 6):** explicit checkbox—LLM may analyze profile text & generate simulated chats. No consent $\rightarrow$ no processing.
- **Data minimization:** only pass necessary profile fields to prompts. No external data sources.
- **Sensitive attributes:** do not infer or reference ethnicity, health, religion, politics, sexual life, etc., unless explicitly volunteered *and* consented for matching; even then, be sparing.
- **Minors:** hard block (KYC-light + age gate).

- **Bias controls:** monitor score distributions across age/language/region; periodic fairness audits.
 - **Storage:** encrypt PII at rest/in transit; pseudonymize prompt/output logs; implement deletion (Right to be Forgotten).
 - **Abuse prevention:** spam/scam classifier, duplicate-profile detection, rate limits.
-

8) Architecture (Prototype)

Services

- **Match Service (API):** inputs candidate pairs, runs filters & scoring, calls LLM.
- **LLM Service:** stateless; prompt templates, safety wrappers, moderation.
- **Embeddings Service:** precompute & cache per profile (Redis).
- **Moderation:** text classifiers (toxicity, sexual content, PII leakage).
- **Feature Store:** embeddings, tags, signals (e.g., PostgreSQL + pgvector).
- **Telemetry:** event capture for A/B and calibration.

Sequence

1. Candidate selection (geo/preferences).
 2. Pre-LLM score → threshold (e.g., ≥ 0.55) triggers conversation simulation.
 3. LLM conversation → LLM summary → moderation → display (2) & (3).
-

9) UX Flow (Pre-Match Card)

- **Card** shows:
 - **Generous Short Summary** (90–140 words)
 - **Compatibility 1–10** + brief “why” (3 signals)
 - Buttons: “Start Chat”, “More Suggestions”
 - Toggle “Show details” → reveals **signals/evidence** (theme overlaps), **not** the full simulated chat.
-

10) Edge Cases & Fallbacks

- **Very short profiles:** “Lite” conversation (≤ 6 turns) + summary focused on questions, not claims.

- **Value conflicts:** phrase kindly (“different perspectives on ...”); score reflects distance.
- **No shared language:** scoring penalty; optional suggestion for simple bridges only if user opted in.
- **Dealbreaker hit:** suppress candidate or show neutral “currently not compatible” (product decision).
- **Moderation failure:** “Preview unavailable” + move to next candidate.

11) Mini Pseudocode

```
def pre_match(a, b):
    if not eligible(a, b):
        return None

    feats = compute_features(a, b) # embeddings, distance, language, lifestyle,
dealbreakers

    raw = weighted_score(feats)

    if raw < LLM_THRESHOLD:
        score = map_to_1_10(calibrate(raw))
        return {"summary": fallback_summary(a, b),
                "score": score, "confidence": "low"}

    convo = llm_simulate(a, b)      # System + User prompts
    summary = llm_summarize(convo, a, b) # Generous + evidence
    if violates_policy(convo) or violates_policy(summary):
        score = map_to_1_10(calibrate(raw))
        return {"summary": safe_fallback(a, b),
                "score": score, "confidence": "low"}

    score = final_score(raw, convo_signals(convo))
```

```
conf = confidence_from_variance(raw, feats)

return {"summary": summary, "score": score, "confidence": conf}
```

12) Test Plan (Prototype)

- **Golden set:** 50 synthetic + 50 real (opt-in) profile pairs.
- **Smoke tests:** determinism (seed/temperature), length bounds, rule violations.
- **Red-teaming:** provocative content, sensitive attributes, phishing-like phrasing.
- **Human review:** 2 reviewers rate Helpful/Warm/Realistic (Likert 1–5).