Dynamic Partner Evaluation Framework

A Conversational, User-Driven Approach to Multi-Dimensional Partner Assessment

---

Executive Summary

PartnerScope's evaluation system enables startups to assess potential partners through a **dynamic, conversational workflow**. Unlike static evaluation frameworks, our approach allows real-time strategy refinement, interactive dimension weighting, and post-evaluation adjustments—all through natural language chat.

Key Features: - Conversational strategy planning - User-configurable evaluation dimensions - Real-time weight adjustment - Post-evaluation refinement (exclude, reweight, filter) - Batch processing for large candidate sets

---

Comparison with PartnerMAS

Our evaluation system is inspired by PartnerMAS (arXiv:2509.24046), but differs significantly in architecture and interaction model.

PartnerMAS (Inspiration)

PartnerMAS uses a **three-tier hierarchical multi-agent system**:

Tier	Agent	Role
1	**Planner Agent**	Designs evaluation strategy
2	**Specialized Agents**	Role-specific assessments (multiple agents in parallel)
3	**Supervisor Agent**	Aggregates outputs into final ranking

Characteristics: - Multiple LLM agents running in parallel - Fixed evaluation strategy (designed once by planner) - Automated aggregation (no human-in-the-loop) - Batch processing of candidates - 10-15% higher match rates vs. single-agent approaches

PartnerScope (Our Implementation)

PartnerScope uses a **single LLM with conversational interface**:

Phase	Process	User Interaction
1	**Strategy Planning**	User can modify dimensions and weights
2	**Evaluation Execution**	Batch processing with progress feedback
3	**Results & Refinement**	User can exclude, reweight, filter results

Characteristics: - Single LLM (not multi-agent) - Dynamic strategy (user can modify at any time) - Human-in-the-loop at every stage - Conversational refinement after evaluation - Interactive, iterative workflow

Key Differences

Aspect	PartnerMAS	PartnerScope
Architecture	Multi-agent hierarchy	Single LLM + chat
Strategy	Fixed (set once)	Dynamic (user-modifiable)
User Role	Passive (receives results)	Active (guides process)
Refinement	Re-run entire pipeline	In-place adjustments
Interaction	Batch input/output	Conversational
Parallelism	Multiple agents in parallel	Sequential batch processing

What we kept from PartnerMAS: - Multi-dimensional evaluation concept - Weighted scoring across dimensions - Strategy-driven assessment

What we changed: - Replaced multi-agent with conversational single-agent - Added dynamic strategy modification - Added post-evaluation refinement - Made the process interactive and iterative

Evaluation Workflow

Phase 1: Strategy Planning

The system proposes an evaluation strategy based on startup profile and candidates.

Process: 1. User initiates evaluation ("start") 2. LLM analyzes startup profile and candidate pool 3. Proposes 4-5 relevant dimensions with weights 4. User can accept or modify

Example Strategy Proposal:

```
1. Market Compatibility (25%)
   Alignment with target market and customer segments

2. Technical Synergy (25%)
   Technology compatibility and integration potential

3. Strategic Alignment (20%)
   Business goals and long-term vision alignment

4. Growth Potential (15%)
   Mutual growth and scalability opportunity

5. Risk Profile (15%)
   Risk factors and potential challenges
```

User Modifications: - "Focus more on technical synergy" - "Add geographic coverage as a dimension" - "Remove risk profile, we'll assess that separately"

The LLM adjusts weights and dimensions accordingly, always ensuring weights sum to 1.0.

Phase 2: Evaluation Execution

Candidates are evaluated in batches to handle large candidate pools efficiently.

Process: 1. Candidates split into batches of 5 2. Each batch evaluated against all dimensions 3. Scores assigned with confidence levels 4. Results aggregated and ranked

Per-Candidate Output:

```json
{
  "candidate_name": "TechPartner Inc",
  "final_score": 85,
  "dimension_scores": [
    {"dimension": "market_compatibility", "score": 90,
        "confidence": 0.85},
    {"dimension": "technical_synergy", "score": 88, "confidence":
        0.90},
    {"dimension": "strategic_alignment", "score": 82,
        "confidence": 0.75}
  ],
  "strengths": ["Strong API ecosystem", "Proven startup
        partnerships"],
  "weaknesses": ["Limited APAC presence"],
```

```
    "recommendations": ["Explore joint pilot program"]
}
```

Scoring Formula:

```
final_score = Σ (dimension_score × weight × confidence) / Σ
(weight × confidence)
```

The confidence factor ensures that dimensions with uncertain data contribute less to the final score.

Phase 3: Results & Refinement

Unlike static evaluation systems, PartnerScope allows **post-evaluation refinement** through natural language.

Refinement Actions:

Action	Example Command	Effect
Exclude	"Remove TechPartner from results"	Filters out candidate, re-ranks remaining
Reweight	"Focus more on technical synergy"	Adjusts weights, recalculates scores
Filter	"Show only top 3"	Returns subset of results
Focus	"Tell me more about geographic coverage"	Provides deeper analysis

Example Refinement Flow:

```
User: "The top candidate has regulatory issues we discovered.
Exclude them."
System: "Removed TechPartner Inc. Re-ranked remaining candidates.
        New #1: DataSync Corp (Score: 82)"

User: "Actually, prioritize geographic coverage more heavily."
System: "Adjusted weights: Geographic Coverage now 30% (was 15%).
        Re-calculated scores. New ranking:
        1. GlobalReach Ltd (Score: 88)
        2. DataSync Corp (Score: 79)"
```

This iterative refinement is **not possible in PartnerMAS**, which requires re-running the entire multi-agent pipeline for any strategy change.
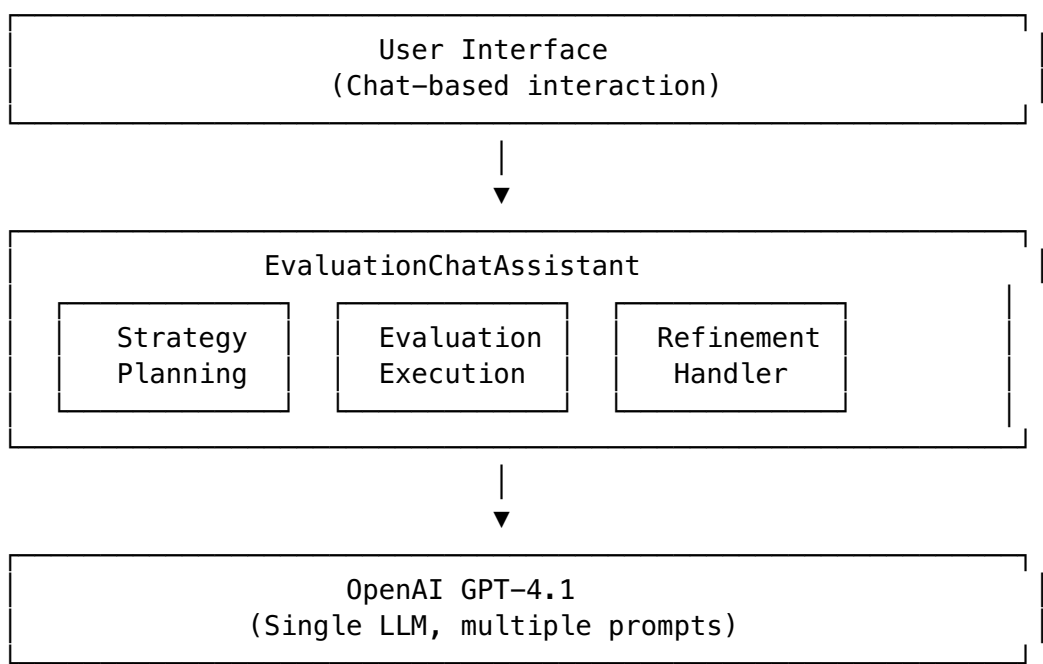

Available Evaluation Dimensions

Dimension	Key Factors
market_compatibility	Market alignment, customer segments, positioning
financial_health	Financial stability, revenue, funding status
technical_synergy	Technology compatibility, integration potential
operational_capacity	Supply chain, logistics, operational capabilities
geographic_coverage	Geographic presence, regional expertise
strategic_alignment	Business goals, long-term vision alignment
cultural_fit	Organizational culture compatibility
resource_complementarity	Complementary resources and expertise
growth_potential	Mutual growth, scalability potential
risk_profile	Risk factors, potential challenges

Users can select any subset of these dimensions and assign custom weights.

Technical Implementation

Architecture

```
┌─────────────────────────────────────────────────────────┐
│                    User Interface                        │
│               (Chat-based interaction)                   │
└─────────────────────────────────────────────────────────┘
                            │
                            ▼
┌─────────────────────────────────────────────────────────┐
│                 EvaluationChatAssistant                  │
│   ┌──────────────┐  ┌──────────────┐  ┌──────────────┐   │
│   │   Strategy   │  │  Evaluation  │  │  Refinement  │   │
│   │   Planning   │  │  Execution   │  │   Handler    │   │
│   └──────────────┘  └──────────────┘  └──────────────┘   │
└─────────────────────────────────────────────────────────┘
                            │
                            ▼
┌─────────────────────────────────────────────────────────┐
│                    OpenAI GPT-4.1                        │
│               (Single LLM, multiple prompts)             │
└─────────────────────────────────────────────────────────┘
```

NOT Multi-Agent

Important clarification: Despite being inspired by PartnerMAS's multi-agent approach, PartnerScope uses a **single LLM** with different prompts for different tasks. This is simpler and enables the conversational, interactive workflow.

PartnerMAS	PartnerScope
Planner Agent	Strategy planning prompt
Specialized Agents	Batch evaluation prompt
Supervisor Agent	Summary generation prompt

We achieve similar task decomposition through **prompt engineering**, not agent orchestration.

Batch Processing

To handle large candidate pools (50+ candidates), evaluation is performed in batches.

Configuration: - Batch size: 5 candidates - Each batch: 1 LLM call - Output: Scores + strengths/weaknesses for each candidate

Why batches? - Avoids context length limits - Maintains evaluation quality (no "context rot") - Enables progress feedback to user - Allows partial results if interrupted

Process:

```
50 candidates → 10 batches → 10 LLM calls → Aggregated ranking
```

Cost Analysis

Operation	LLM Calls	Est. Cost
Strategy proposal	1	$0.02
Strategy modification	1 per change	$0.02
Evaluation (50 candidates)	10 batches + 1 summary	$0.25
Refinement	1 per request	$0.02
Typical session	**~15 calls**	**$0.30-0.50**

Advantages of Dynamic Approach

1. Adaptability

Users can adjust strategy mid-process based on: - New information about candidates - Changing business priorities - Stakeholder feedback

2. Transparency

Every decision is visible and explainable: - Dimension weights are explicit - Scores include confidence levels - Reasoning is provided for rankings

3. Iteration Speed

Refinements don't require re-evaluation: - Exclude a candidate: instant re-rank - Adjust weights: instant recalculation - Full re-evaluation only when needed

4. Human Judgment

The system augments, not replaces, human decision-making: - User controls strategy - User validates results - User applies domain knowledge

---

Limitations

vs. PartnerMAS

Limitation	Explanation
No parallel processing	Single LLM, sequential batches
No specialized expertise	Same model for all dimensions
Potential inconsistency	Different batches may have slight scoring variation

Mitigations

- **Batch processing** reduces context rot
- **Confidence scores** flag uncertain evaluations
- **User refinement** allows correction of anomalies

Example Session

```
User: Start evaluation

System: Here's my proposed evaluation strategy:
```

1. Market Compatibility (25%)
        2. Technical Synergy (25%)
        3. Strategic Alignment (20%)
        4. Growth Potential (15%)
        5. Risk Profile (15%)

        This will evaluate 23 candidates. Adjust weights or run?

User: Focus more on technical synergy, we need API integration

System: Adjusted weights:
        — Technical Synergy: 35% (was 25%)
        — Market Compatibility: 20% (was 25%)

        Ready to run evaluation?

User: Yes, run it

System: Evaluating 23 candidates in 5 batches...
        Batch 1/5 complete ✓
        Batch 2/5 complete ✓
        ...

        ## Evaluation Complete!

        1. APIConnect Inc — Score: 92
           + Strong REST API, developer-friendly docs
        2. TechBridge Ltd — Score: 87
           + Proven integrations with 50+ platforms
        3. DataSync Corp — Score: 84
           + Real-time sync capabilities

User: Exclude APIConnect, they're a competitor

System: Removed APIConnect Inc. Re-ranked:
        1. TechBridge Ltd — Score: 87
        2. DataSync Corp — Score: 84
        3. CloudLink Partners — Score: 81

User: Tell me more about TechBridge's weaknesses

System: TechBridge Ltd weaknesses:
        — Limited presence in APAC region
        — Smaller support team (response times 24–48h)
        — No SOC2 certification yet (in progress)

_____

References

1. "PartnerMAS: An LLM Hierarchical Multi-Agent Framework for Business Partner Selection on High-Dimensional Features." arXiv:2509.24046v1, 2025. https://arxiv.org/abs/2509.24046
2. OpenAI. "GPT-4.1 Technical Report." 2025.

Document Version: 1.0 Last Updated: January 2026 PartnerScope by UTokyo Research