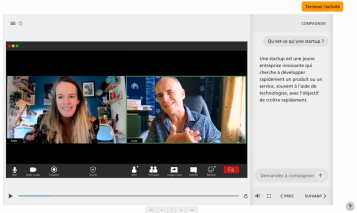
Can AI Mentoring Build Durable Skills?

Testing Mentor Filtering Capacity at Scale



Marcos Balmaceda

KU Leuven — *2026-2027 Job Market*

with

Thomas Åstebro
HEC Paris

Bruno Crépon
ENSAE-CREST

Mona Mensmann
Univ. Cologne

Mathis Schulte
HEC Paris

Andrew Funck
HEC Paris

Naja Pape
INSEAD

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

UCL School of Management — AI Plus Management Doctoral Consortium, 2026

The Challenge: Scaling What Actually Works

Entrepreneurial Soft-Skills Training

Quality: It works

- Personal initiative: **+30%** profits
  Campos et al. (2017)
- Negotiation: **+4 p.p.** enrolment
  Ashraf et al. (2020)

Scale: Requires expert facilitators

- Online: <10% MOOC completion
  Claflin et al. (2021)
- Low gains despite incentives
  Asanov et al. (2023)

Bastani et al. (2025): AI tutoring with guardrails → guardrails mitigate learning loss after AI removal

Poulidis et al. (2025): "self-regulated" AI use (= unrestricted help-seeking) → −34 p.p. learning
chess RCT, $n = 216$

We build on — SRL (Jin et al., 2023):
Phase (3) × Area (4) × Strategy (10)
Conditions the chatbot response to the learner's current cognitive and motivational state.

The Challenge: Scaling What Actually Works

What remains to be understood

The field has not yet studied how the **role** of a chatbot as a learning companion shapes skill formation.

Theory: The Economics of Mentoring

Mentorship vs. Neighboring Roles

Role	Dominant dimension	Horizon
Mentorship	Filtering (σ_m) + full relational bundle	Long-term
Coaching	Decision architecture (ΔD)	Short-term
Teaching	Knowledge (ΔK)	Episodic

Pignataro (2026) defines coaching and teaching as limiting cases of mentorship, where most dimensions collapse to zero. Mentorship's distinctive value is filtering—the dimension they cannot replicate.

Theory: The Economics of Mentoring

Mentoring is formalized through **success entropy** $H_{p,t}$ — the learner's uncertainty about which actions lead to desirable outcomes. Its dynamics:

$$\dot{H}_{p,t} = -\sum_{m \in M} q_{p,m}(t)\, \sigma_m H_{p,t} + \eta - \delta_H H_{p,t}$$

$q_{p,m}$ = interaction intensity σ_m = mentor filtering capacity η = environmental noise δ_H = natural entropy decay

Our focus: holding all other dimensions constant by design, we vary σ_m directly — does the mentor's filtering capacity, embedded in an AI chatbot, produce skills that persist after the course ends?

Theoretical Predictions

	Filtering Mentor	Open Mentor
During the course	Fast gains	Slower gains
After course ends	Decays	**Persists**
Why	AI Mentor selected one angle from the materials	AI mentor kept multiple angles open

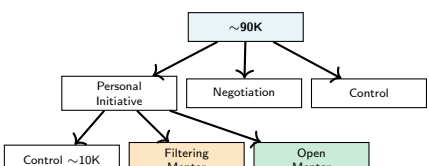
Which matters depends on the policy objective: immediate performance vs. durable skill formation.
Chatbot active *only during the course*; measurements during follow-ups are purely observational.

Setting & Experimental Design

France Travail
- ∼**90,000** job seekers enrolled
- 12–18 month follow-up

Courses (mental models, not business practices)
- **Personal Initiative** Campos et al., 2017
- **Negotiation Skills** Ashraf et al., 2020

Stratified Cross-Randomization



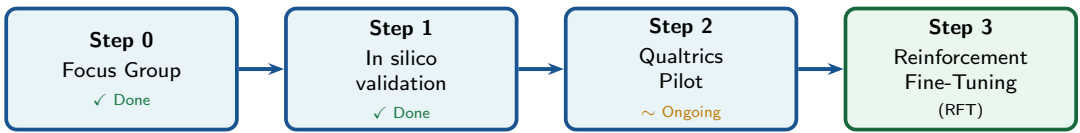
Three Conditions

Control: No chatbot — standard content only

Filtering Mentor (high σ_m)
"Identify the ONE most relevant angle. One clear path. No alternatives."
`L: ''I'm not sure this training is useful''`
`M: ''What sign would let you judge it useful?''`

Open Mentor (low σ_m)
"Do not select a single frame. Open possibilities and let them navigate."
`L: ''I'm not sure this training is useful''`
`M: ''Several angles --- what matters most now?''`

Reinforcing σ_m Separation: From Prompts to Model Weights



- GPT-5.4 API + RFT pipeline (via OpenAI Research Access)
- Select golden examples (top 100)

Output: Two fine-tuned chatbots — making treatment separation **robust**.

In Silico Validation: Confirming σ_m Separation Before Field Deployment

Framework: GABE

Tranchero et al. (2025) In silico experiments to validate theory **before** costly field deployment.

→ 100 LLM-generated French job-seeker messages, 5 runs per condition across all 9 SRL states.

Metric: Frame Count Score (FCS)

0.0 = one clear path (Filtering Mentor)
1.0 = multiple open framings (Open Mentor)
Threshold: $\text{FCS}_{\text{Open}} - \text{FCS}_{\text{Filtering}} > 0.20$

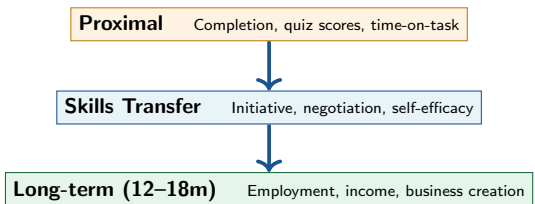
Model Selection Results

Model	Filt.	Open	Gap	
gpt-4o-mini (orig.)	0.247	0.293	+0.046	△
gpt-4o-mini (rev.)	0.287	0.403	+0.116	△
gpt-5.4, low	**0.117**	**0.473**	**+0.356**	✓
gpt-5.5, low	0.086	0.457	+0.371	×

△ Below threshold ✓ Validated × Latency >4s (confound)

Contributions & Feedback

Outcome Cascade



1. **Mechanism:** Does selecting your own angle produce skills that last?
2. **Scale:** Can mentoring reach millions via AI?
3. **Equity:** AI removes the selection barrier to elite mentoring.
4. **External validity:** France (unemployed adults) → LatAm (students entering the labour market). Same mechanism?

Critical test: Do skills built through self-navigation persist at 12–18 months *after* the course ends?

Contact
marcos.balmaceda@kuleuven.be