INVITED PLENARY · 14TH INTERNATIONAL SCIENTIFIC CONFERENCE “COMPUTER SCIENCE” · SOZOPOL, 13–15 SEPTEMBER 2026

From Stakeholder Needs to Deployed Practice

Designing and evaluating a governed generative-AI assistant for entrepreneurship education

Angel Marinov
Technical University of Varna · a.marinov@tu-varna.bg

with H. Nenov, O. Gungor, R. Montoliu, F. Alniak, S. Franch-Ventura, M. Ivanovici, C. Bungeag, L. Staršelskis, D. Vasilev, E. Mihaylov

TWO PAPERS, ONE STORY

A · From stakeholder needs to institutional deployment

B · Two-stage evaluation: expert and user evidence across five functional areas

Erasmus+ SEEN · BG01-KA220-HED-000255491

Co-funded by the European Union

Mentoring does not scale. Unguided AI already does.

86%

of surveyed students already use AI tools such as ChatGPT

26%

have received any formal or informal AI training

THE INSTITUTIONAL QUESTION

What does the system know — and who decided?

Can its knowledge be inspected and corrected? What happens to user data? Who pays, and what does it cost at scale?

1 Individualised feedback is the scarcest resource

Entrepreneurial learning needs timely, iterative feedback on the learner's own idea. Human mentoring does not scale with enrolment.

2 Public chatbots answer from an opaque knowledge base

No control over sources, pedagogy, data protection or cost. The institution cannot audit, align or defend the tool.

3 The gap is institutional, not algorithmic

The question is not whether LLMs can help, but whether a university can offer AI support it governs.

Co-funded by the European Union

Eight partners, five countries, one work package on AI



Consortium at Transilvania University of Brașov, 2025

Technical University of Varna
Bulgaria

Innovation Education Lab
Romania

Bulgarian Business Forums
Bulgaria

Dare to Develop Academy
Bulgaria

Efektas Group
Lithuania

PUHU
Netherlands

Transilvania University of Brașov
Romania

Universitat Jaume I
Spain

Erasmus+ KA220-HED

Cooperation partnership in higher education, Sept 2024 – Aug 2026

Bulgaria · Spain · Romania · Netherlands · Lithuania

Universities plus ecosystem partners: NGOs, business forums, an AI R&D company

WP3 · Generative AI training and deployment

Design-based sequence: needs → architecture → knowledge base & deployment → expert test → field test

Co-funded by the European Union

Evidence first: no build decision preceded the survey

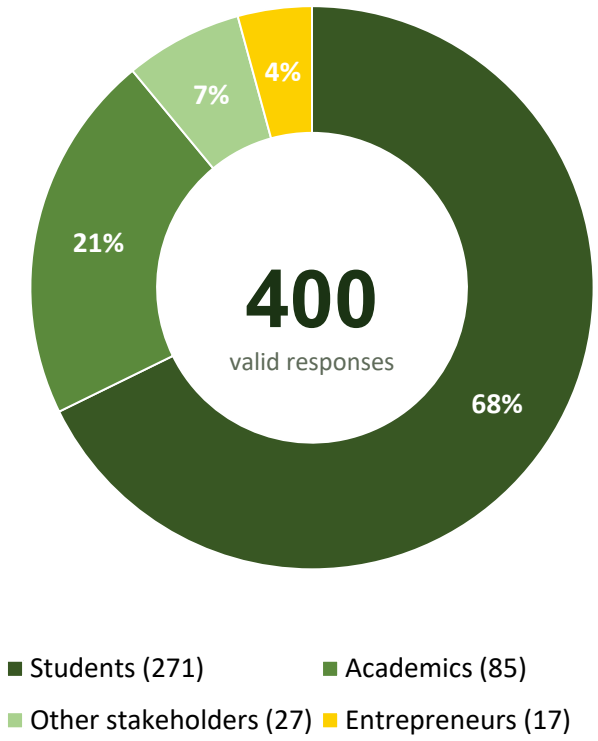


Each stage produced the input of the next.

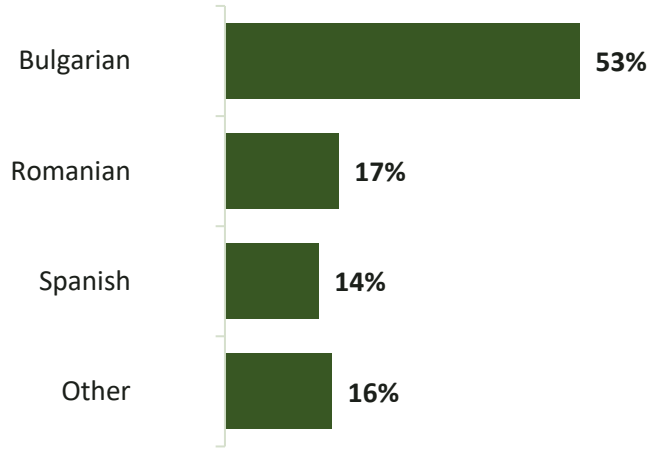
The needs analysis fixed the functional requirements; the architecture comparison fixed the implementation; the functional specification then became the acceptance reference and, later, the evaluation instrument. This ordering is what distinguishes the project from technology-first pilots that procure a tool and look for use cases afterwards.

Co-funded by the European Union

Four languages, four countries, 400 stakeholders



Nationality of respondents



Convenience sampling through partner channels; the coordinating Bulgarian university is over-represented. Results are descriptive, not nationally representative.

WHAT WAS ASKED

Students

Career intentions · support received vs desired · importance of 4 entrepreneurship and 6 AI topics · AI use, usefulness, training · interest in institutional AI tools

Academics

Same topic and AI blocks with educator wording · willingness and incentives to support student ventures

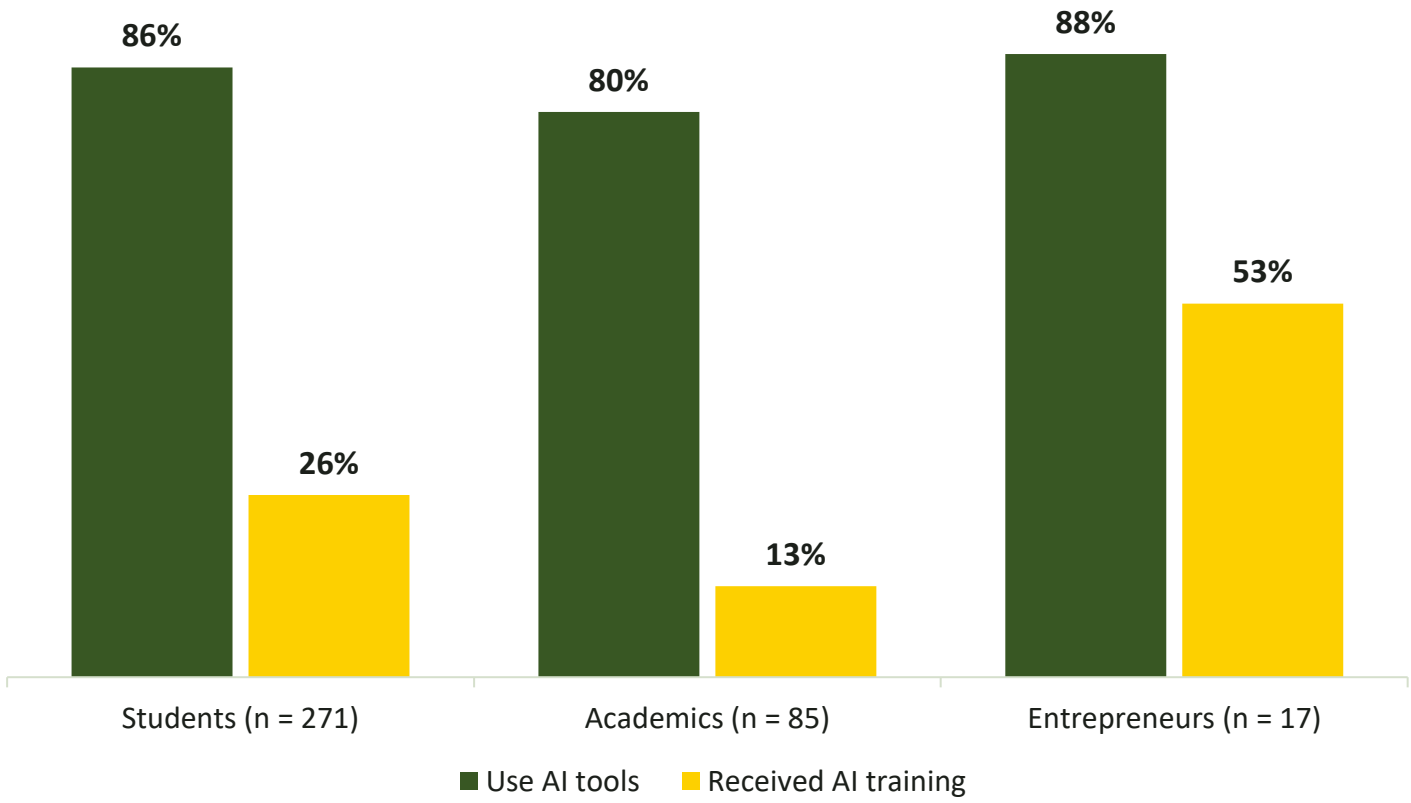
Entrepreneurs & stakeholders

AI in business operations · importance of AI tools and student innovation in higher education

Languages: BG · EN · RO · ES. Instrument public: github.com/astmarinov/SEEN-Surveys

Co-funded by the European Union

Adoption runs far ahead of preparation in every group



Perceived usefulness among student users

3.77 / 5

Academic use counts any of teaching, research or development. A further 33% of academics work in computer science and were treated as domain-prepared.

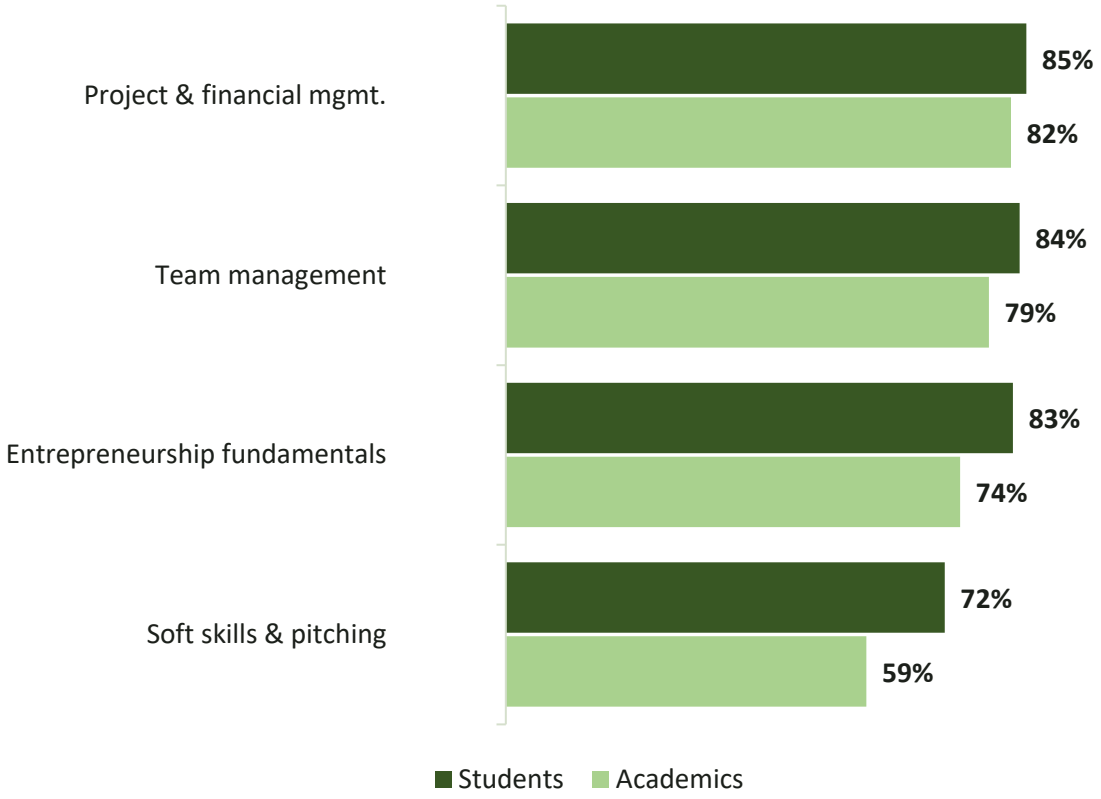
WHY IT MATTERS FOR DESIGN

A governed assistant responds to an existing, largely unguided practice — it does not create a new one. Guidance must therefore be applied rather than introductory, and verifiable rather than oracular.

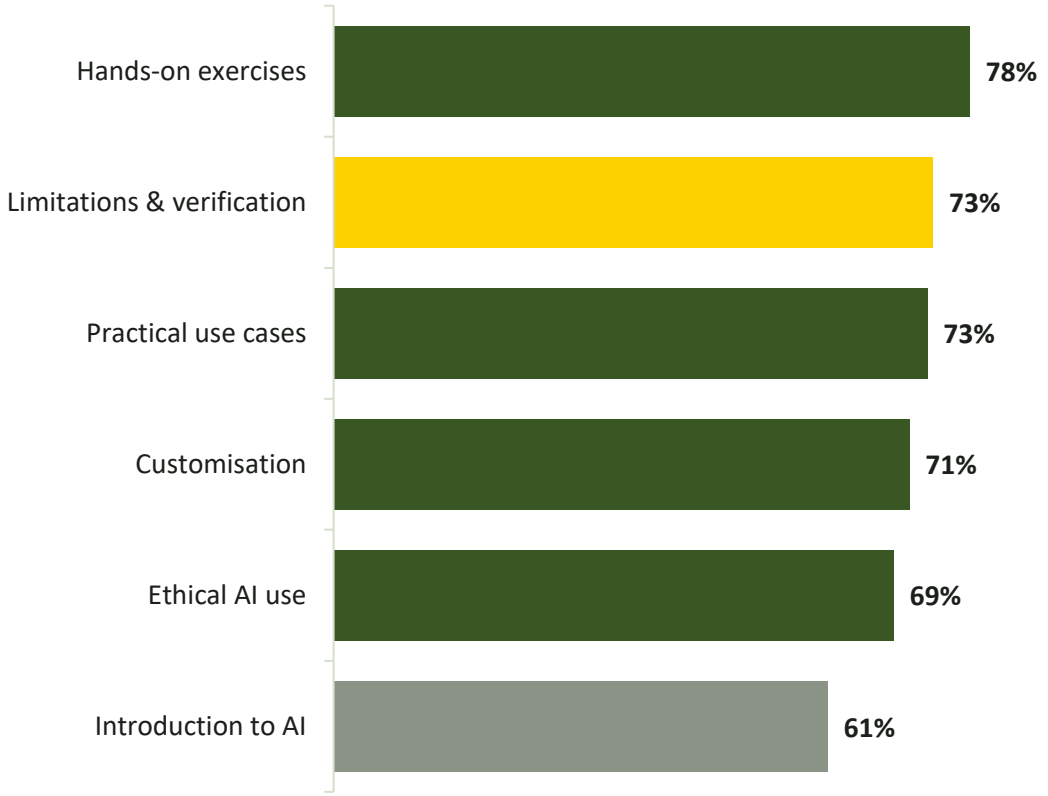
Co-funded by the European Union

The demand is for applied, critical AI literacy

Entrepreneurship course topics rated important or very important



AI course topics rated important or very important (students)



Limitations & verification outrank the introduction by 13 points — the demand for verifiable outputs is explicit.

Co-funded by the European Union

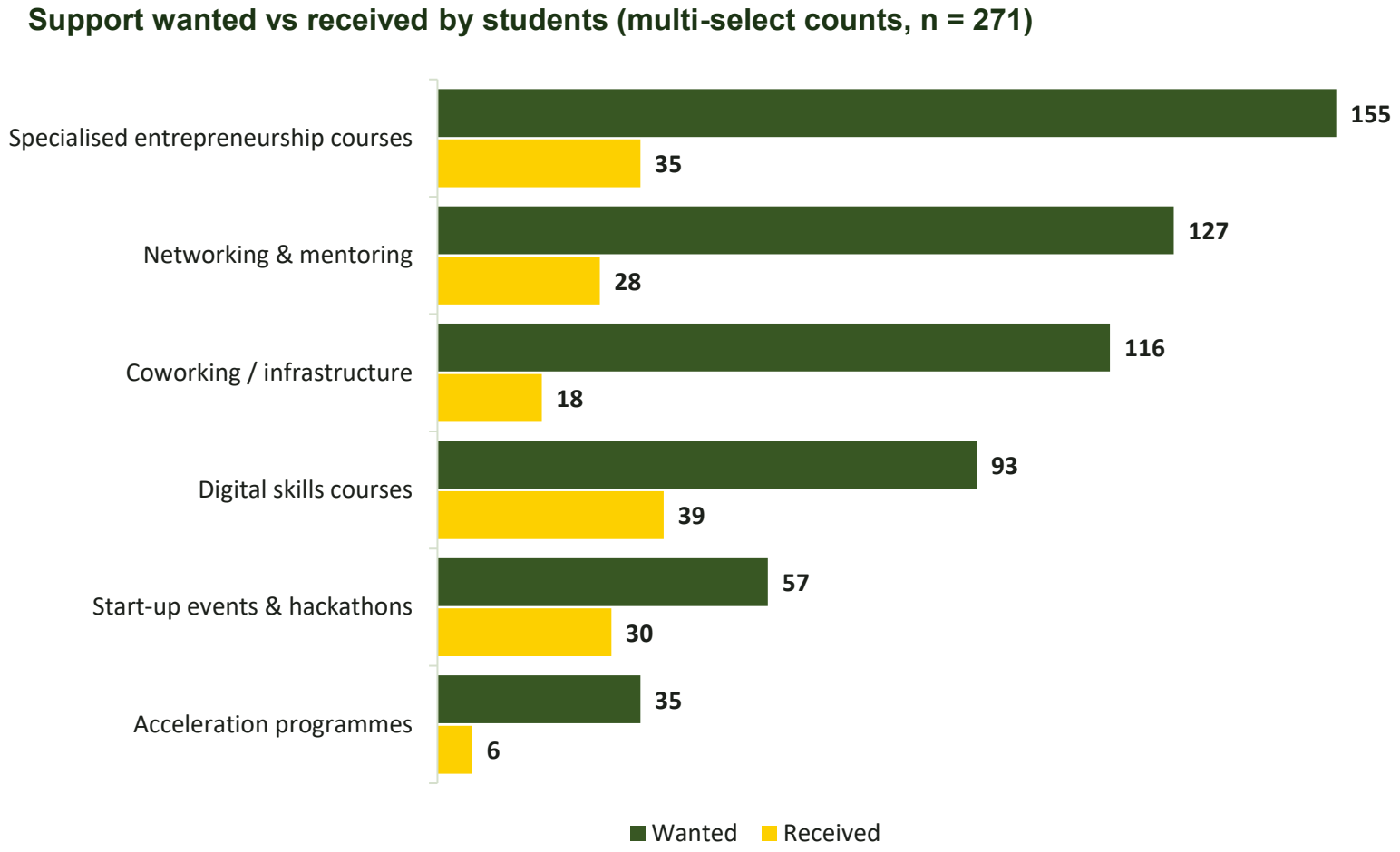
Two-thirds of students get no entrepreneurship support

182 / 271

students reported no support at all relevant to starting a business

THE GAP IS WIDEST WHERE HUMAN ATTENTION IS REQUIRED

Courses, mentoring and coworking are the most wanted and the most attention-intensive forms of support. A tool giving individualised, on-demand guidance at negligible marginal cost addresses exactly this gap.



Co-funded by the European Union

They want the institution to provide the tool

55%

of students strongly interested (4–5 / 5) in custom AI tools developed by their own institution

M = 3.55 / 5

60%

of academics strongly interested in institution-developed AI tools

M = 3.82 / 5

4.19

mean importance of AI tools in education, rated by other higher-education stakeholders (1–5)

n = 27

THREE FINDINGS DEFINE THE REQUIREMENTS SPACE

Individualised, on-demand guidance addresses the widest support gaps · institutional governance matches the preference for custom tools and the demand for verifiable outputs · applied, critically informed interaction meets the literacy demand.

Co-funded by the European Union

NEED ANALYSIS → REQUIREMENTS

Every design decision traces back to a survey finding

EVIDENCE	REQUIREMENT	SYSTEM RESPONSE
High AI use, little formal training	**Applied guidance with verifiable sources**	**Grounded responses with source provenance**
Two-thirds of students without institutional support	**On-demand, scalable individualised support**	**Five structured entrepreneurship workflows**
Strong interest in institution-provided tools	**University governance and identity**	**Authenticated deployment on institutional infrastructure**
Demand for practical and critical AI literacy	**Structured interaction with visible limitations**	**Stage-specific workflows, constrained prompting, provenance**
Erasmus+ funding ends in 2026	**Inspectable, updatable knowledge base**	**Curated corpus, administrator upload, incremental indexing**

Co-funded by the European Union

Three approaches, five requirement-derived criteria

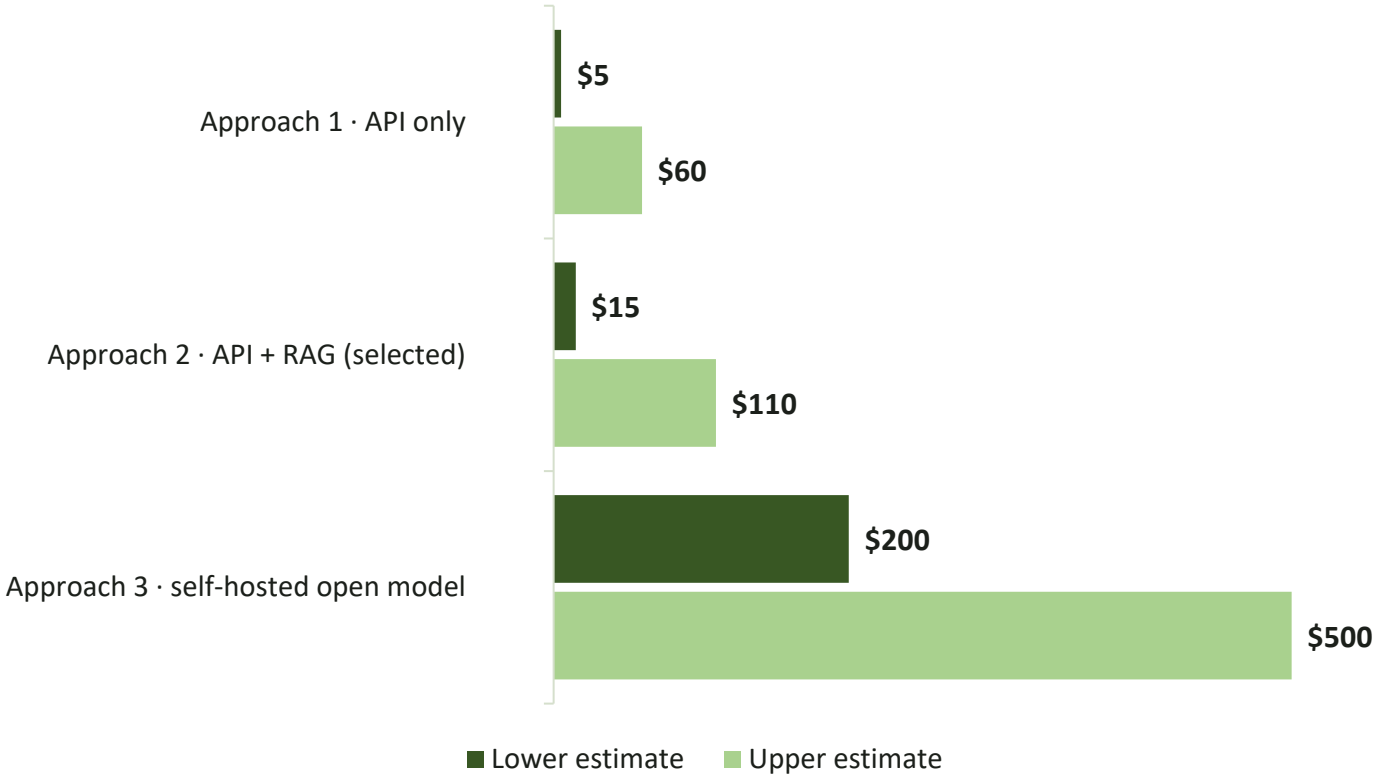
	APPROACH 1 API only	APPROACH 2 · SELECTED API + RAG over a curated corpus	APPROACH 3 Custom / self-hosted open model
Domain specificity	Low — generic model knowledge	**High — grounded in curated corpus**	High — if trained on domain data
Content governance & provenance	None	**Full — inspectable corpus, cited passages**	Full
Data protection	Provider-dependent	Provider-dependent; no special-category data	Maximal — on-premises
Implementation & maintenance effort	Minimal	Moderate — corpus pipeline, vector index	High — training, MLOps, hardware
Indicative monthly cost	≈ USD 5–60	≈ USD 15–110	≈ USD 200–500 flat

Assessment at time of selection (2025); costs indicative for 300–1000 requests/month. A design decision against project needs, not a controlled benchmark.

Co-funded by the European Union

Tens of dollars a month, not a data centre

Indicative monthly operating cost, USD (selection-time estimates, 2025)



Ranges span 300–1000 requests/month and the shortlisted providers; self-hosting is a flat infrastructure cost regardless of volume.

MODELS SHORTLISTED IN THE 2025 STUDY

GPT-4 · Claude 3 · Gemini Pro · Command R+ · Amazon Titan · Mistral 7B, with Pinecone / LangChain-style retrieval stacks for Approach 2 and LLaMA, Mixtral, Falcon, Flan-T5 for self-hosting.

Usage scenarios: 100 students × 3 or × 10 interactions per month (300 / 1000 requests).

CAVEAT

Per-token prices have fallen sharply since selection. That strengthens rather than alters the ordering: what is durable is that a governed deployment is within an ordinary university unit's budget.

Co-funded by the European Union

RAG is an institutional choice, not a technical fashion

PUBLIC CHATBOT		GOVERNED RAG DEPLOYMENT
Opaque, shifting knowledge base	→	**Curated corpus the institution can inspect and update**
Cannot cite institutional materials	→	**Responses carry provenance links to vetted passages**
No role-based control	→	**Authenticated access, RBAC, admin panel**
Data logged under external terms	→	**Minimal, GDPR-aware logging bound to local policy**
Can change or disappear without notice	→	**Service evolves only when the institution decides**
Needs ML specialists to adapt	→	**Maintained by ordinary academic staff: upload a document**

None of this requires the underlying model to be better than the public one. The value lies in the wrapper of governance around it.

Co-funded by
the European Union

DEPLOYED SYSTEM · ARCHITECTURE

Four cooperating layers, wrapped in operational safeguards

Interaction layer
Authenticated web UI · five structured workflows · role-based admin panel (users, corpus) · secure PDF/TXT upload

Orchestration layer
Multi-step agentic flow with dedicated agent roles · input validation · retrieval · grounded prompt assembly around task templates (problem–solution fit, market hypotheses, value proposition, risks, next steps)

Knowledge layer (RAG)
Curated corpus (PDF + TXT) · transformer embeddings · managed vector index (Pinecone) with provenance metadata

Generation layer
Model-agnostic API access under institutional credentials (currently OpenAI GPT-4 family) · scope, tone and grounding constraints · admin-configurable agent behaviour

Operational safeguards
Role-based access control · GDPR-aware minimal logging · secret management via environment variables · cost guardrails at the API boundary · response provenance for audit

FIVE STRUCTURED WORKFLOWS

1. **Idea validation** — problem–solution fit, validation planning
2. **Business model** — structure and revenue logic
3. **Next steps** — prioritised action planning
4. **Resources** — matching to learning resources
5. **Pitch evaluation** — structured feedback on pitch texts

The workflows mirror canonical early-venture stages — and they also define the five functional areas of the evaluation.

Co-funded by the European Union

Growing the corpus is an administrative task

1

Curate

Academic & practitioner papers, business-model and lean-startup handbooks, project-management guidance, AI-adoption reports — screened for licensing, topical fit and absence of personal data

2

Segment

Long documents split into passages that preserve local coherence; short TXT items carry rubrics, glossaries, checklists

3

Embed

Transformer-based embedding model maps each passage to a vector (create_embeddings.py)

4

Index

Managed vector index retains minimal provenance: source path, file type (index upload script)

5

Extend

Partners upload vetted documents through the admin panel; incremental re-indexing, no redeployment

WHAT THE CORPUS IS GOOD AT

Frameworks and structured analysis: business-model canvases, lean-startup cycles, validation methods, project planning — knowledge that already lives in documents.

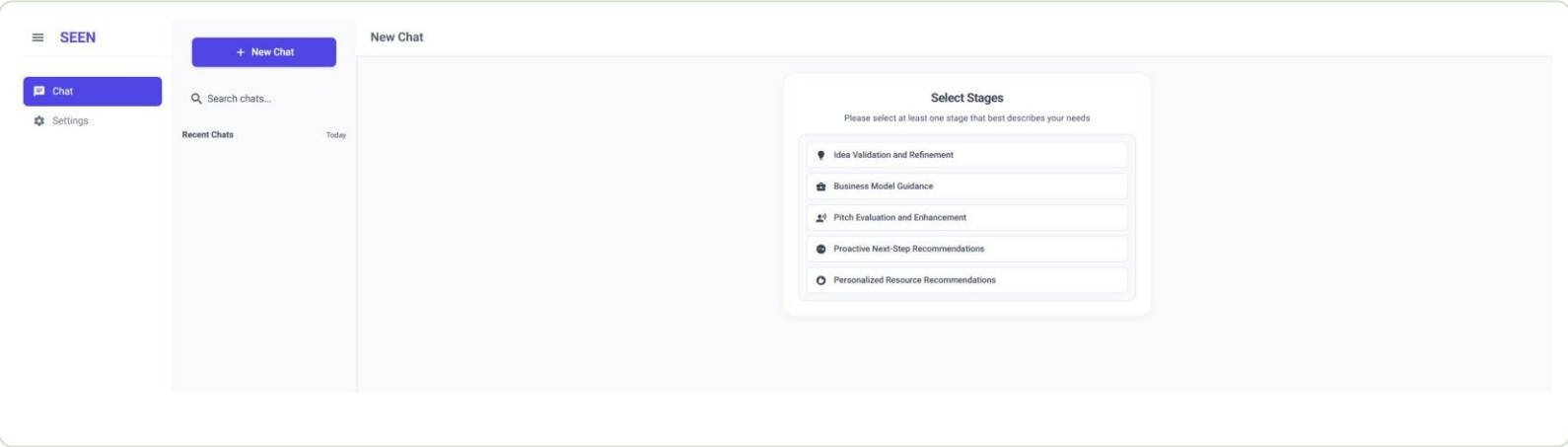
WHAT IT IS NATURALLY POOR AT (remember this)

Knowledge that lives in practice rather than in print — for instance graded examples of weak and strong pitches with the transformations between them.

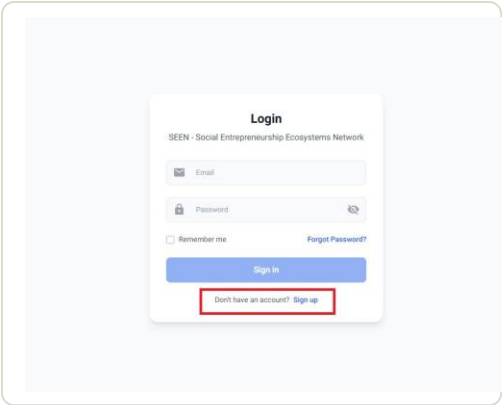
Co-funded by the European Union

DEPLOYED SYSTEM · THE PRODUCT

In production on university servers: seen.tu-varna.bg



Stage selector: the user picks the workflow before the first message — scope is part of the user experience.



DEPLOYMENT PRACTICE

- HTTPS on the institution's official subdomain, behind a reverse proxy
- Role-based access control; institutional credentials at the API boundary
- Minimal, purpose-bound logging; no special-category data; removal on request
- Cost guardrails at the API boundary; secrets via environment variables
- Admin panel: users, roles, corpus upload, index refresh, agent configuration

WHAT A USER GETS

Grounded feedback

a human-readable brief plus a structured object: strengths, gaps, recommendations, referenced sources

Provenance

each response links back to the retrieved passages, so a sceptical student or instructor can check where the advice comes from

Visible limits

stage-specific workflows and a defined no-answer path when the corpus does not contain the evidence

Export for mentor handover

structured feedback can be taken to a human mentor as a first pass

Co-funded by the European Union

Ethics by design: every principle paired with a control

1 Transparency & traceability

Visible citations; minimal source metadata per passage; transparency pack (model card, data statement, change log)

2 Grounded by default

Retrieval-first prompting; uncertainty cues; a defined no-answer / needs-human-review path

3 Scope control

On-topic system prompts; input pre-processing limits; authenticated access; role-based corpus updates

4 Fairness monitoring

Rotating suite of standard queries across founder profiles, sectors, languages; corpus diversity log; rebalancing

5 User rights & redress

In-interface flag-and-review channel with time-bound resolution and traceable notes

6 Privacy & security by design

Purpose-bound minimal logs; no special-category data; removal on request; secrets in environment variables

Codified in the WP3 good-practice report before feedback collection began, and grounded in the survey's own demand for ethics, verification and transparency.

Co-funded by the European Union

Built incrementally; every stage closed on a verified outcome



FOUR PARALLEL ACCEPTANCE STREAMS

- Stakeholder scenario testing against a functional-scenario spreadsheet owned by the academic team — re-executed after every significant change
- Automated backend test suite
- Access-control review against stakeholder feedback
- End-to-end operational verification from a normal user browser after each production-side change

ROBUSTNESS IN PRODUCTION

Three defect classes surfaced in acceptance testing: user-facing behaviour not matching documented scenarios · an admin permission restriction · a cross-origin / port configuration inherited from development.

None was structural. The architecture selected at the requirements stage proved sufficient under real institutional conditions.

Caveat: absence of observed failure at pilot scale, not a demonstration of scalability under high concurrent load.

Co-funded by the European Union

Two independent stages: expert test, then field test

STAGE 1 · EXPERT TEST

18
raters — consortium staff and affiliated experts

50
predefined scenarios, ten per functional area

7 of 15
criteria per scenario, fixed in the instrument

350
scenario–criterion combinations; 2–16 ratings each

Inputs held constant: every rater judges the same scenarios, so variation reflects the system, not the prompt. Criterion-level diagnosis; nothing about authentic use.

STAGE 2 · FIELD TEST

114
users after hands-on use; 94% students

15 + 6
functional-area items and platform-experience items

α 0.81–0.93
internal consistency of the five area scales

647
substantive free-text entries, inductively coded

Users bring their own ideas: authenticity at the cost of standardisation. Same five-area structure, no shared raters, no shared materials.

Co-funded by the European Union

PILOTING · SCALE OF USE

Real use at pilot scale in three universities

151
registered platform users

472
chats

4,124
messages exchanged

2,710
AI requests served

1.39 M
tokens processed





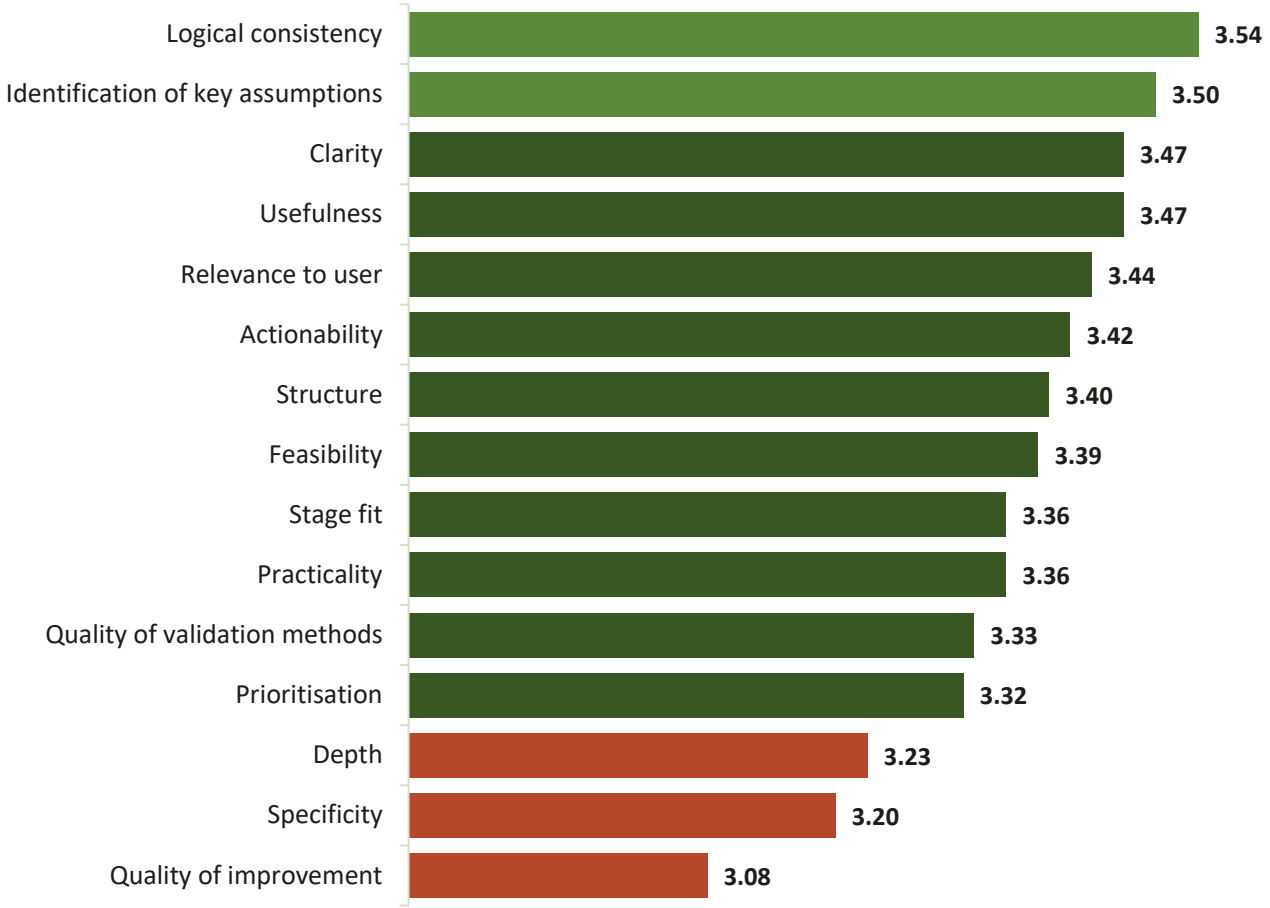
WHO TESTED

Students at Transilvania University of Brașov, Universitat Jaume I and Technical University of Varna, in courses and project events, April–June 2026.

Usage totals ≠ evaluation sample: 151 used the platform, 114 completed the evaluation questionnaire, never linked to accounts.

Co-funded by the European Union

Experts: strong on reasoning, weak on tailoring



Unweighted average of the 350 scenario–criterion combination means; axis starts at 2.8 to show structure.

3.37

overall expert mean on the five-point scale — above the midpoint, well below the ceiling

STRENGTHS · organisation and reasoning

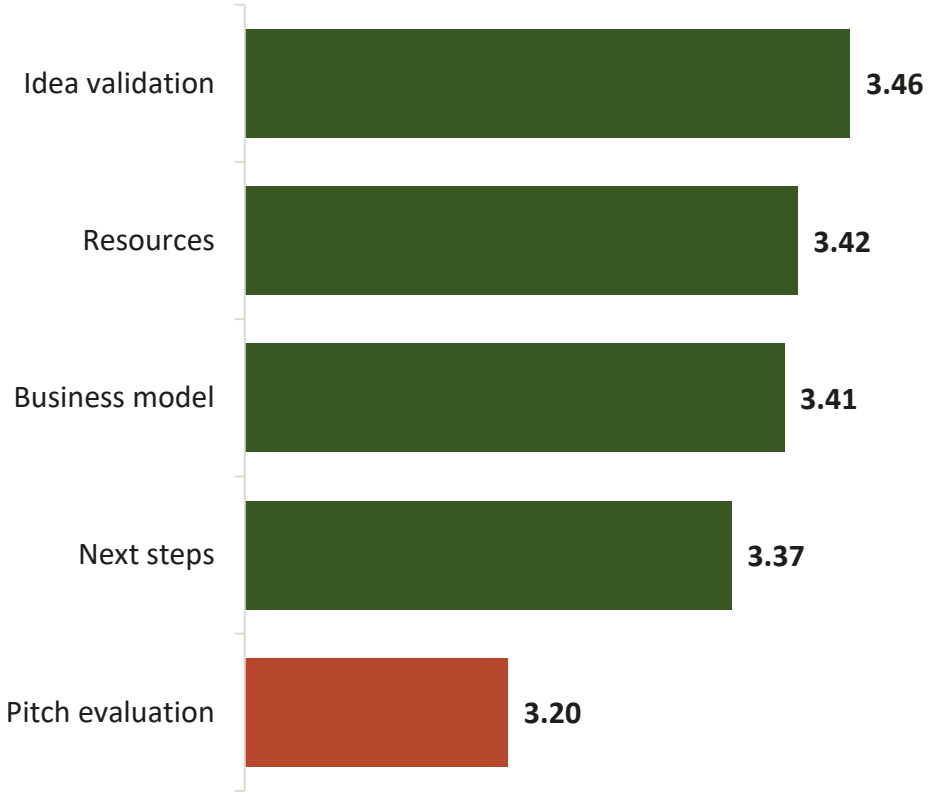
Logical consistency 3.54 · key assumptions 3.50 · clarity 3.47 · usefulness 3.47

WEAKNESSES · tailoring and transformation

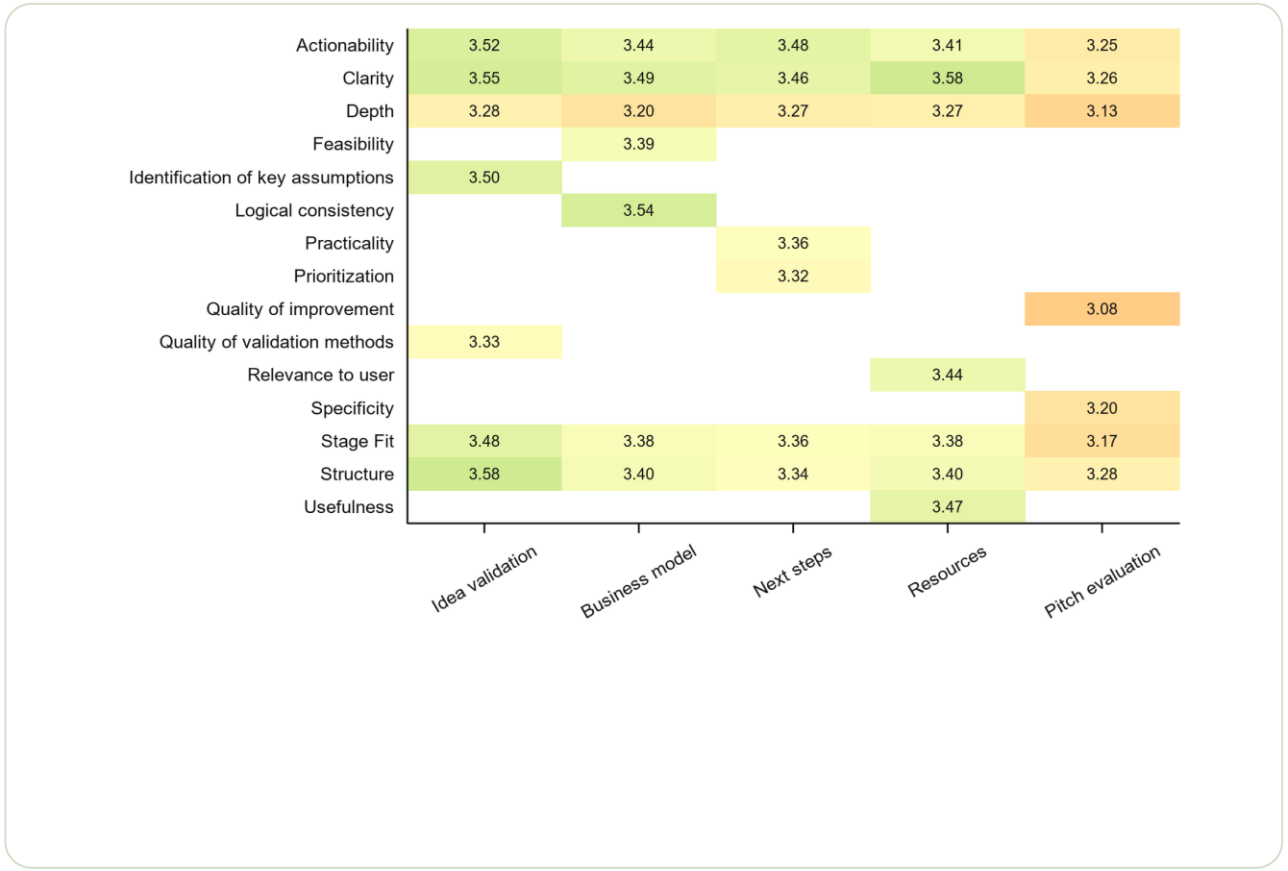
Quality of improvement 3.08 · specificity 3.20 · depth 3.23

Co-funded by the European Union

By area, pitch evaluation is clearly the weakest



Instrument-defined areas: every scenario belonged to exactly one block of ten, fixed before data collection.



Mean rating by criterion × area; blank cells = criterion not in that area's rubric. The pitch column is the lowest on every shared criterion.

Co-funded by the European Union

The weakest scenarios ask to transform, not to explain

“Turn into elevator pitch: A service for sharing student resources.”	**2.61**
“Rewrite this pitch: We created a website where students can sell their notes.”	**2.92**
“I am building an online course platform for students. Should I use subscriptions or one-time payments?”	**2.95**
“I validated my idea with 10 students. What should I do next?”	**3.04**
“What business model works for a student marketplace for freelance services?”	**3.05**
“Make this pitch clearer: We help students find jobs faster using AI.”	**3.08**

Six lowest-rated of 50 scenarios (expert mean). Shaded = pitch-rewriting tasks.

THE COMMON DENOMINATOR

Explanation tasks
organise, structure, explain a framework — the system is rated stronger

Transformation tasks
rewrite this pitch, choose between two concrete options, decide what to do after ten interviews — the system is rated weaker

Four of the seven lowest-rated scenarios are pitch rewrites.

Co-funded by the European Union

RESULTS · STAGE 2 FIELD TEST

114 users: pitch evaluation lowest again, significantly

Business model

3.65

Explanation clear	3.79
Model realistic	3.56
Helped structure idea	3.60

Resources

3.62

Recommendations relevant	3.77
Resources useful	3.66
Matched my level	3.44

Next steps

3.60

Steps clear	3.80
Practical, actionable	3.62
Matched my situation	3.38

Idea validation

3.50

Understand viability	3.58
Identified risks	3.30
Practical, useful	3.63

Pitch evaluation

3.32

Improved my idea	3.40
Suggestions specific	3.29
Made pitch clearer	3.27

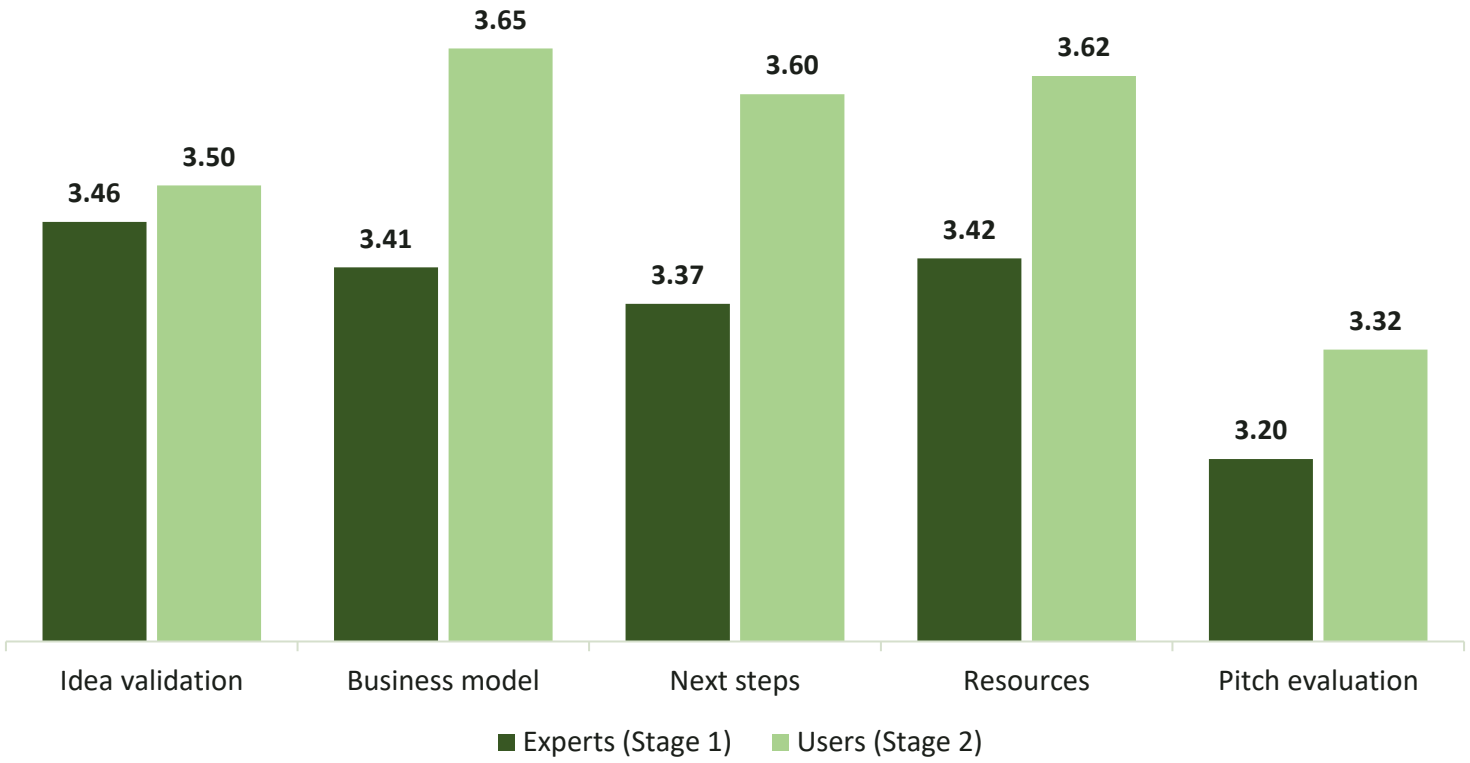
PAIRED COMPARISON · pitch vs mean of the other four areas

3.32 vs 3.59 · t(113) = −3.19 · p = 0.002 · Cohen's d = −0.30

Within every area the concrete, transformation-oriented item trails the explanatory ones — e.g. “identified important risks” 3.30 vs “suggestions practical” 3.63 in idea validation.

Co-funded by the European Union

Users are more lenient, but the rank order is the same



WHY THE AGREEMENT CARRIES WEIGHT

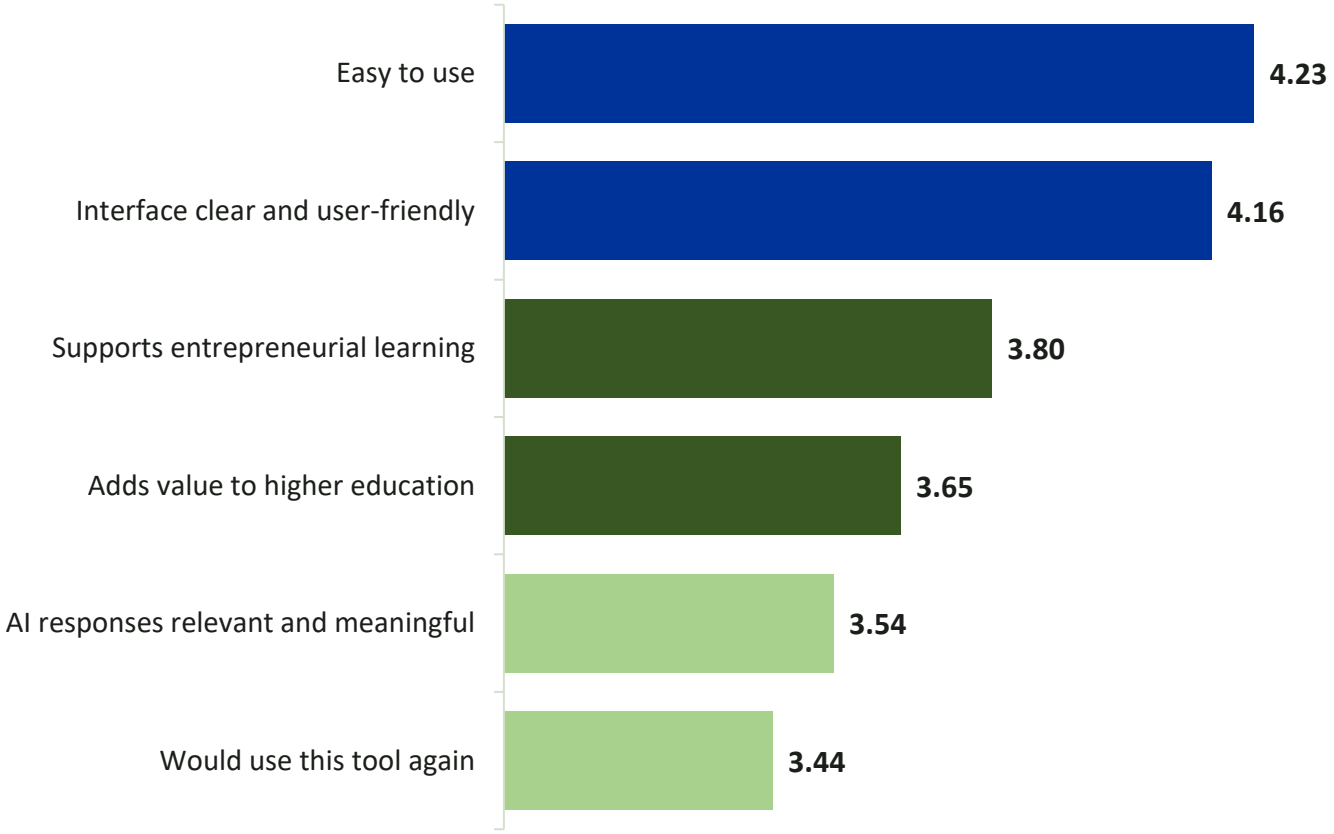
The two stages differ in

- raters — experts vs end users
- materials — fixed scenarios vs own ideas
- instruments — criterion rubrics vs Likert + free text
- setting — controlled vs authentic use

Shared-method bias is therefore an unlikely explanation. What they share is the system — so the pitch weakness is a property of the platform, not an artefact of either design.

Co-funded by
the European Union

Product vs advice: usability is not the bottleneck



Six platform-experience items, n = 114, five-point scales, scale α = 0.89.

ARTEFACT

Ease of use 4.23 and interface clarity 4.16 — the highest ratings in the whole study.

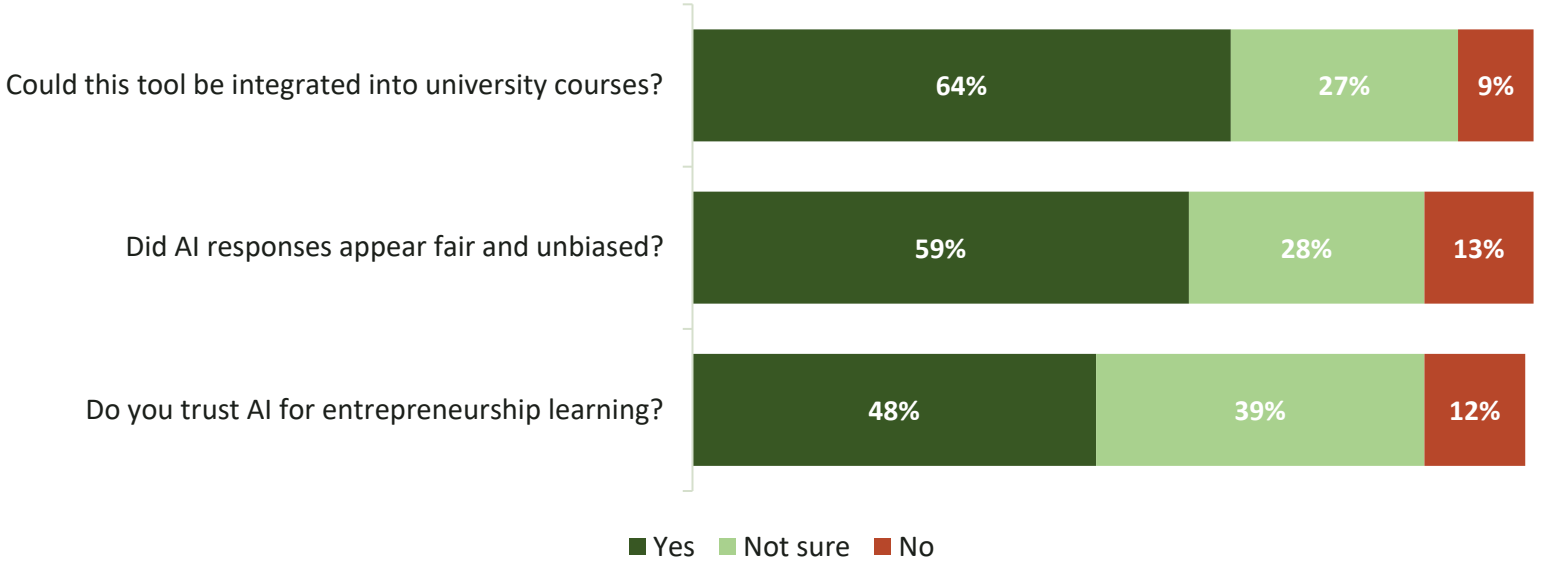
ADVICE

Intention to reuse (3.44) tracks the content-quality ratings (3.32–3.65), not the ease-of-use ratings. Reuse decisions follow perceived advice quality, not interface satisfaction.

→ improvement effort belongs in corpus quality and coverage, not interface polish.

Co-funded by the European Union

Calibrated trust: positive on product, measured on advice



n = 114; percentages of respondents. Counts: 73/31/10 · 67/32/15 · 55/45/14.

HOW THE PLATFORM SUPPORTED LEARNING (multi-select)

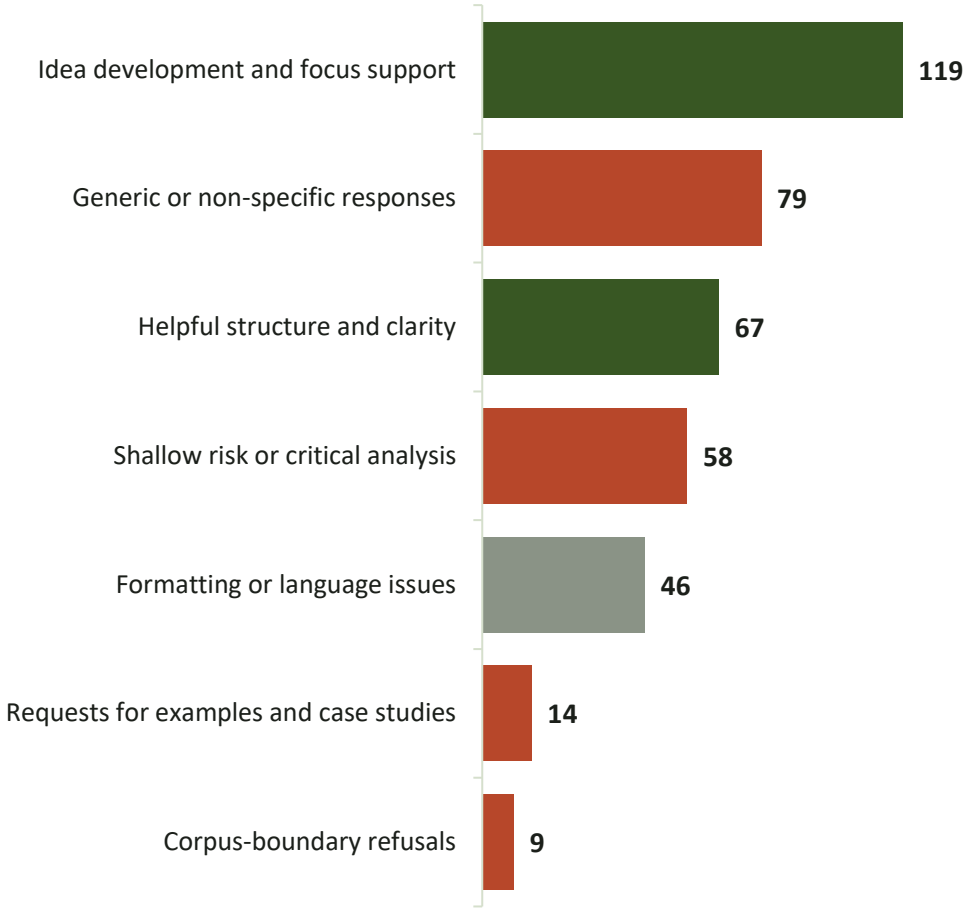
Idea development · 74
Personalised guidance · 55
Skill development · 48
Critical thinking · 37
Innovation mindset · 30

READ THE 39% UNDECIDED AS AN ASSET

Calibrated rather than blind trust is exactly what the human–AI trust literature recommends, and it matches the applied, critically informed AI literacy the needs analysis asked for. The 64% support for course integration is a receptive context — provided integration preserves that scepticism.

Co-funded by the European Union

647 comments: praised for structure, criticised as generic



"It helped me understand the idea in more detail, leaving the ambiguity away, and made me focus on the main issue."

"More innovative or interesting ideas could not really be validated, and the feedback was often too generic."

"It provided me with structured, quite detailed information without me asking for it."

"The overall overview was useful, but I was looking for a more critical analysis of the potential flaws."

"It used raw markdown (like stars) instead of actual bold text, which made it harder to read."

"More advanced customization and real startup case studies would improve the experience."

"Regarding my business idea, the AI said it doesn't have that information."

Coded mentions, not respondent percentages; entries could carry multiple codes. Keyword-assisted inductive coding.

Co-funded by the European Union

RESULTS · CONVERGENT INTERPRETATION

Five observations align across two independent stages

1. **Experts and users** both ranked pitch evaluation lowest of the five areas
2. **Both stages** located the weakness in specificity
3. **Experts** additionally identified limited depth and quality of improvement
4. **Users** described responses as generic and asked for concrete examples
5. **A minority of users** documented refusals at the corpus boundary

THE GENERAL LESSON FOR EDUCATIONAL RAG

Corpus composition is a pedagogical decision. Content types that live in practice rather than in print will be underserved unless deliberately represented.

A PLAUSIBLE MECHANISM

A document-based corpus is rich in frameworks — handbooks embed readily as text — but poor in what pitch feedback needs: graded examples of weak and strong pitches with explicit transformations between them. Pitching is taught through exemplars and live iteration; its knowledge lives in practice, not in print.

NOT A DEMONSTRATED CAUSE

No corpus ablation was run. Weaker pitch prompt templates, a model less capable at constrained rewriting, or response-length limits cannot be excluded. The intervention on the next slide is designed to discriminate.

Co-funded by the European Union

Closing the pitch gap with expert-validated exemplars

1

Rubric anchoring

Pitch-quality rubric derived from the very criteria that diagnosed the gap: specificity, quality of improvement, structure, stage fit

2

Exemplar generation

LLM-generated weak → improved pitch pairs, annotated with the transformation applied (quantify the problem, name the customer, sharpen the differentiator); ~10 sectors × 3 stages × 10 pairs ≈ 300

3

Expert validation

Every pair screened with the Stage-1 instrument; only pairs above thresholds on improvement quality and specificity survive — the evaluation instrument becomes a quality gate for synthetic data

4

Corpus integration

Validated exemplars embedded into the RAG index; the two-stage evaluation is repeated with the existing instruments and aggregation rules

quality feedback loop: evaluation results steer subsequent generation rounds

SUCCESS CRITERION

A rise in the pitch-area mean and in specificity and improvement quality toward the other four areas, measured with the same instruments as the baseline.

IT ALSO DISCRIMINATES

If validated exemplars move the pitch ratings, corpus coverage was the binding constraint. If they do not, attention shifts to prompting and model choice. No architectural change; low marginal cost.

Co-funded by the European Union

LESSONS FOR INSTITUTIONS

Six lessons from design and evaluation

FROM THE DESIGN CHAIN

1 Design from evidence

Every architectural decision traced to a survey finding; the requirements spreadsheet doubled as the acceptance reference and, later, the evaluation instrument.

2 Spend effort on the corpus

The architecture is assembled from mature components; corpus quality, coverage and maintainability are where design effort pays.

3 Governance is affordable

Tens of dollars a month at pilot scale; after funding ends, maintenance is a document collection plus an API subscription.

FROM THE EVALUATION

4 Evaluate functions, not averages

The pitch gap is invisible in a single satisfaction score; it appears only in disaggregated, criterion-level measurement.

5 Usability is not the bottleneck

Ease of use 4.2, content quality 3.3–3.7; improvement effort belongs in corpus quality and coverage, not interface polish.

6 Calibrated trust is an asset

48% trust, 39% undecided: pair the assistant with instructor oversight — assistant first, instructor second — and teach with the tool and about its limits at once.

Co-funded by
the European Union

BEYOND SEEN

The chain generalises beyond entrepreneurship

Legal-clinic education
case files and procedural guidance as corpus

Thesis supervision support
methodological handbooks, faculty guidelines

Engineering design review
standards, checklists, past design reports

Any exemplar-borne domain
…with the synthetic-exemplar pipeline

Substitute the corpus and the workflows; the architecture and the governance model stay the same.

WHAT CONTINUES AFTER AUGUST 2026

- Platform stays on university infrastructure; corpus maintained through the admin panel
- Scenario suite kept as a standing regression instrument against provider-side model changes
- Synthetic pitch-exemplar study and re-evaluation with the same instruments
- Course integration with instructor oversight at partner universities

WHAT WE DO NOT CLAIM

- Deployment shows feasibility, not learning effectiveness — that needs longitudinal designs
- Convenience samples; expert raters were affiliated, not blinded; no generic-chatbot control
- Retrieval mitigates but does not eliminate hallucination; the commercial model is a moving dependency
- One deployment, one corpus, one model — a single case, honestly measured

Co-funded by the European Union

Thank you

Evidence → requirements → governed architecture → controlled deployment → function-level evaluation

Angel Marinov

Technical University of Varna

a.marinov@tu-varna.bg

Papers: From Stakeholder Needs to Institutional Deployment (A) · Two-Stage Evaluation of a RAG Assistant for Entrepreneurship Education (B) — MDPI Engineering Proceedings, Computer Science 2026

SEEN AI platform
seen.tu-varna.bg

All instruments (S1–S3)
github.com/astmarinov/SEEN-Surveys

Co-funded by the European Union through Erasmus+, project "Social Entrepreneurship Ecosystems Network" (SEEN), grant BG01-KA220-HED-000255491. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or EACEA.

Manuscript text refinement used Anthropic's Claude; the authors take full responsibility for the content.

Co-funded by the European Union