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

Two-Stage Evaluation of a RAG Assistant for Entrepreneurship Education

Expert and user evidence across five functional areas

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

IN ONE LINE

18 experts, 50 scenarios and 114 users agree: content quality adequate, usability high, pitch evaluation weakest — and the average would have hidden it.

Erasmus+ SEEN · BG01-KA220-HED-000255491

Co-funded by the European Union

MOTIVATION

Aggregate satisfaction can hide task-level failure

1

Coarse evidence

Evaluations of educational chatbots and LLMs mostly report overall satisfaction or acceptance at whole-system level. A system can be easy to use and well organised while underperforming at the task learners need most.

2

Corpus dependence

A RAG system's output quality is bounded by corpus composition. Content types that live in practice rather than in documents will be underserved — an effect invisible in aggregate ratings and rarely examined.

3

So: evaluate functions

Measure at the level of individual pedagogical functions and quality criteria, with two independent sources of evidence, and look for the weaknesses that recur across both.

THREE RESEARCH QUESTIONS

RQ1 How does the assistant perform against expert quality criteria across five entrepreneurship-support functions?

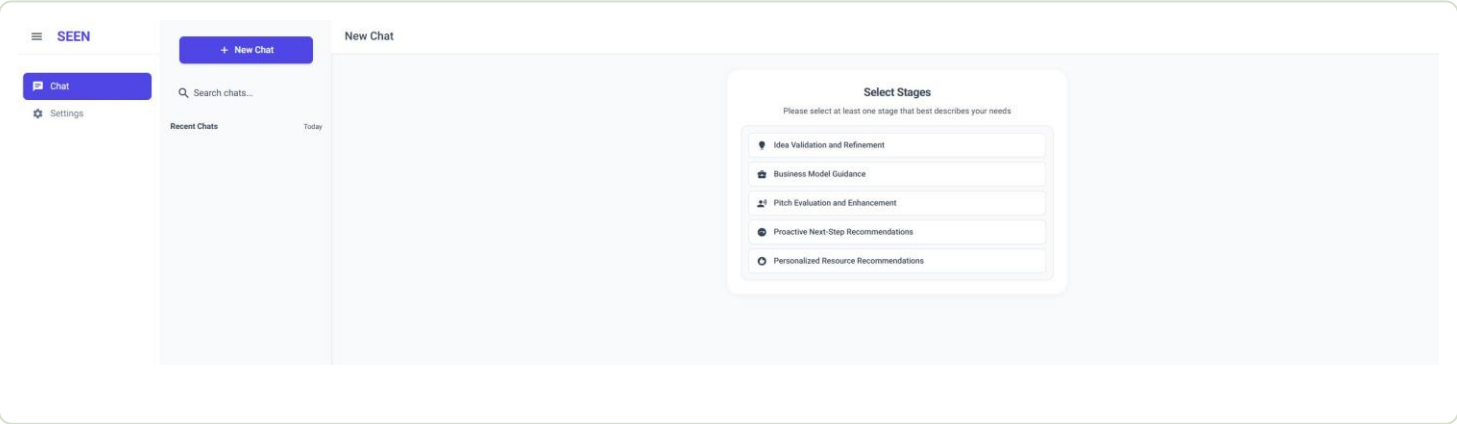
RQ2 How do end users evaluate the five functions, the platform experience, fairness, trust and suitability for course integration?

RQ3 Which function- and criterion-level weaknesses emerge consistently across expert ratings, user ratings and coded feedback?

Co-funded by
the European Union

THE SYSTEM UNDER EVALUATION

A deployed RAG assistant with five structured workflows



seen.tu-varna.bg — the user selects a workflow before the first message.

Idea validation
viability, risks, validation methods

Business model
structure, realism, explanation

Resources
relevant materials matched to level

Pitch evaluation
clarity, specificity, improvement

Next steps
practical, situation-matched actions

WHAT IT IS

- Retrieval-augmented generation over a curated corpus of entrepreneurship handbooks, papers and guidance; commercial LLM via API; provenance links in every answer.
- In production on Technical University of Varna infrastructure; authenticated access; admin-managed corpus.
- Motivated by a four-country needs analysis (n = 400): pervasive but untrained AI use and unmet demand for individualised support.
- The five workflows come from the functional specification — and they define the five areas of this evaluation.

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

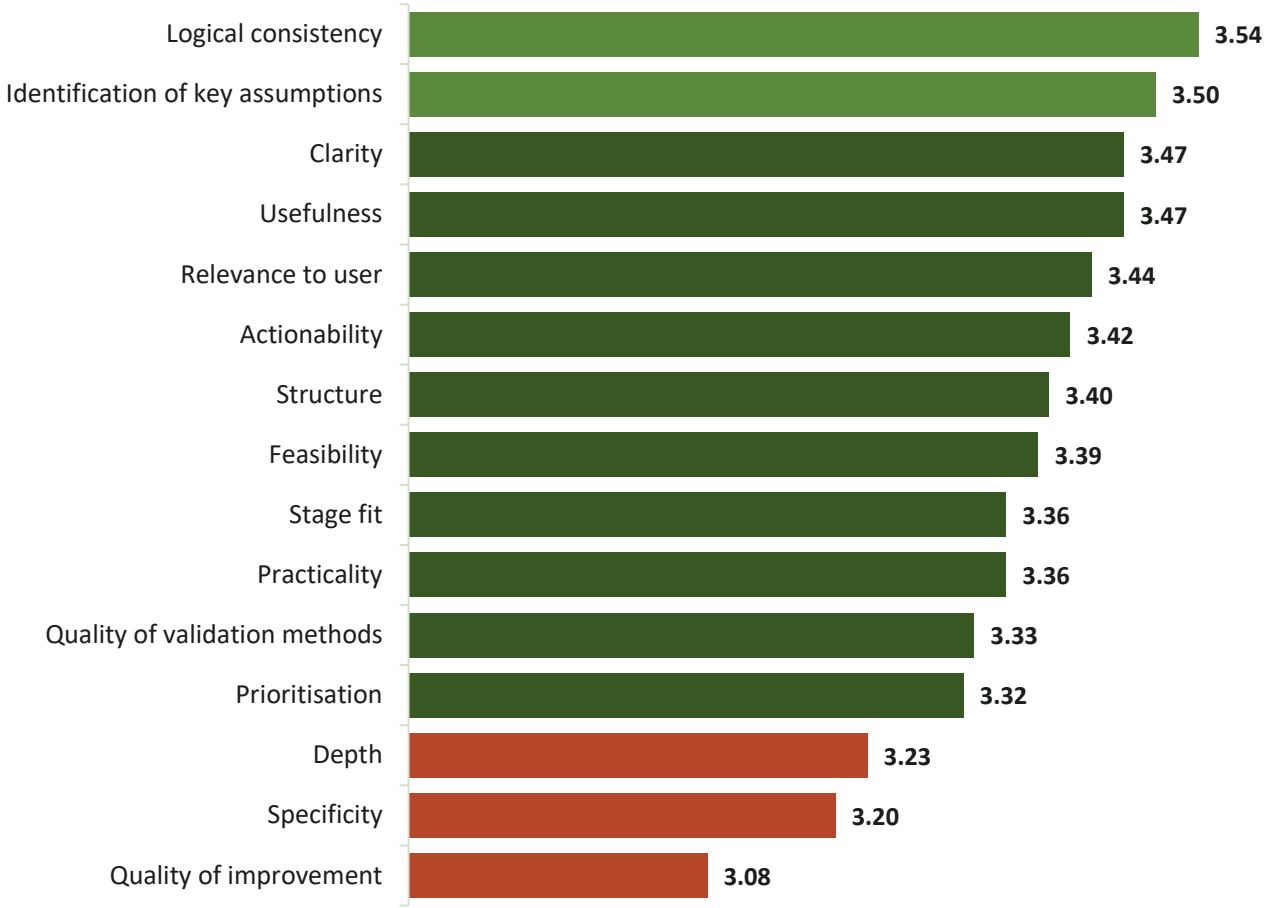
α 0.81–0.93
internal consistency of the five area scales

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

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

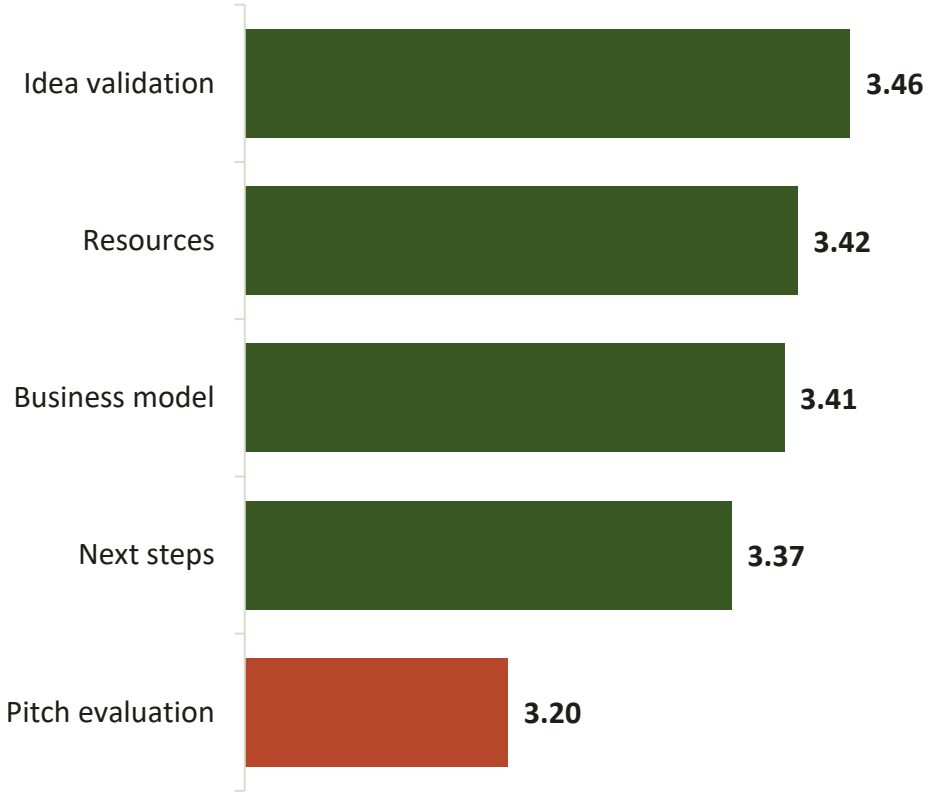
Logical consistency 3.54 · key assumptions 3.50 · clarity 3.47 · usefulness 3.47

WEAKNESSES · tailoring and transformation

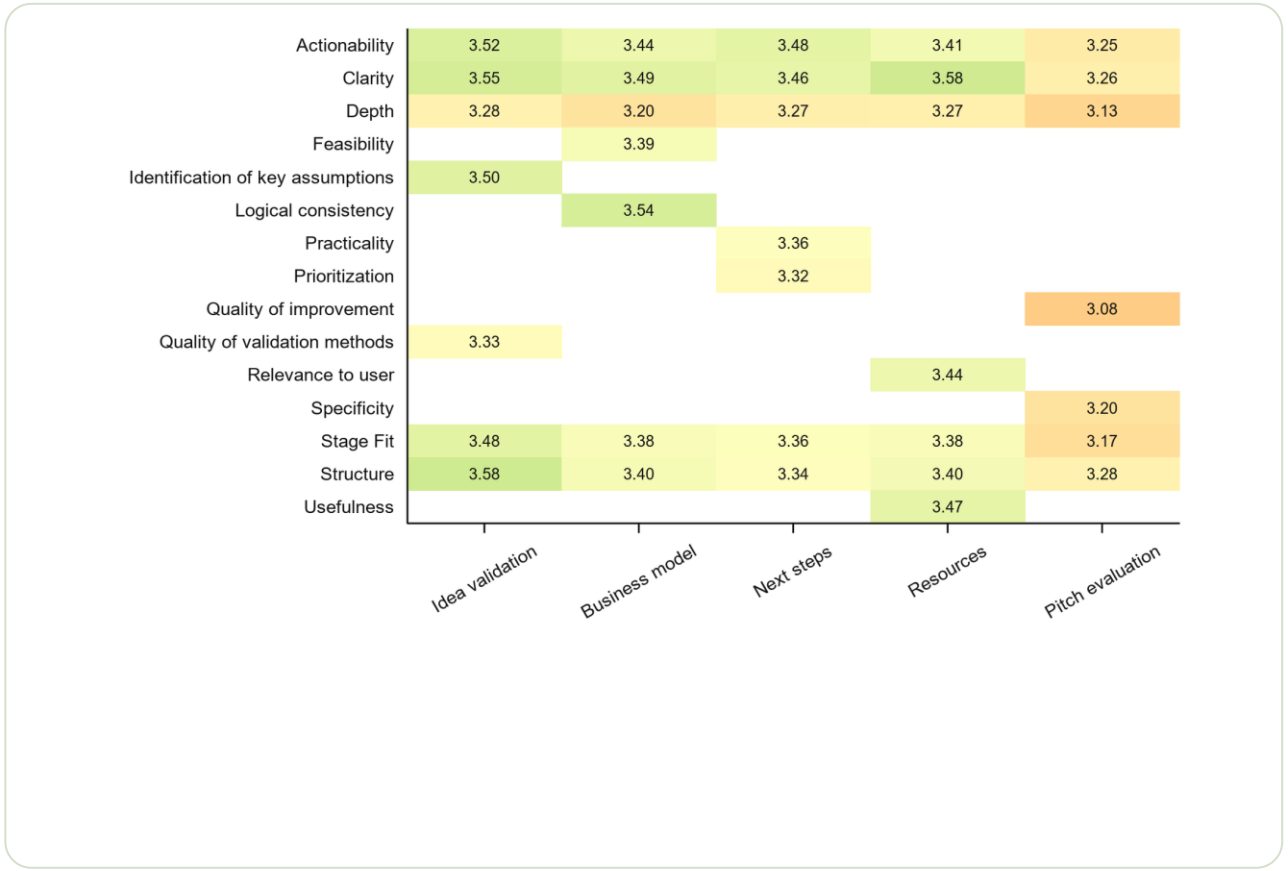
Quality of improvement 3.08 · specificity 3.20 · depth 3.23

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

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

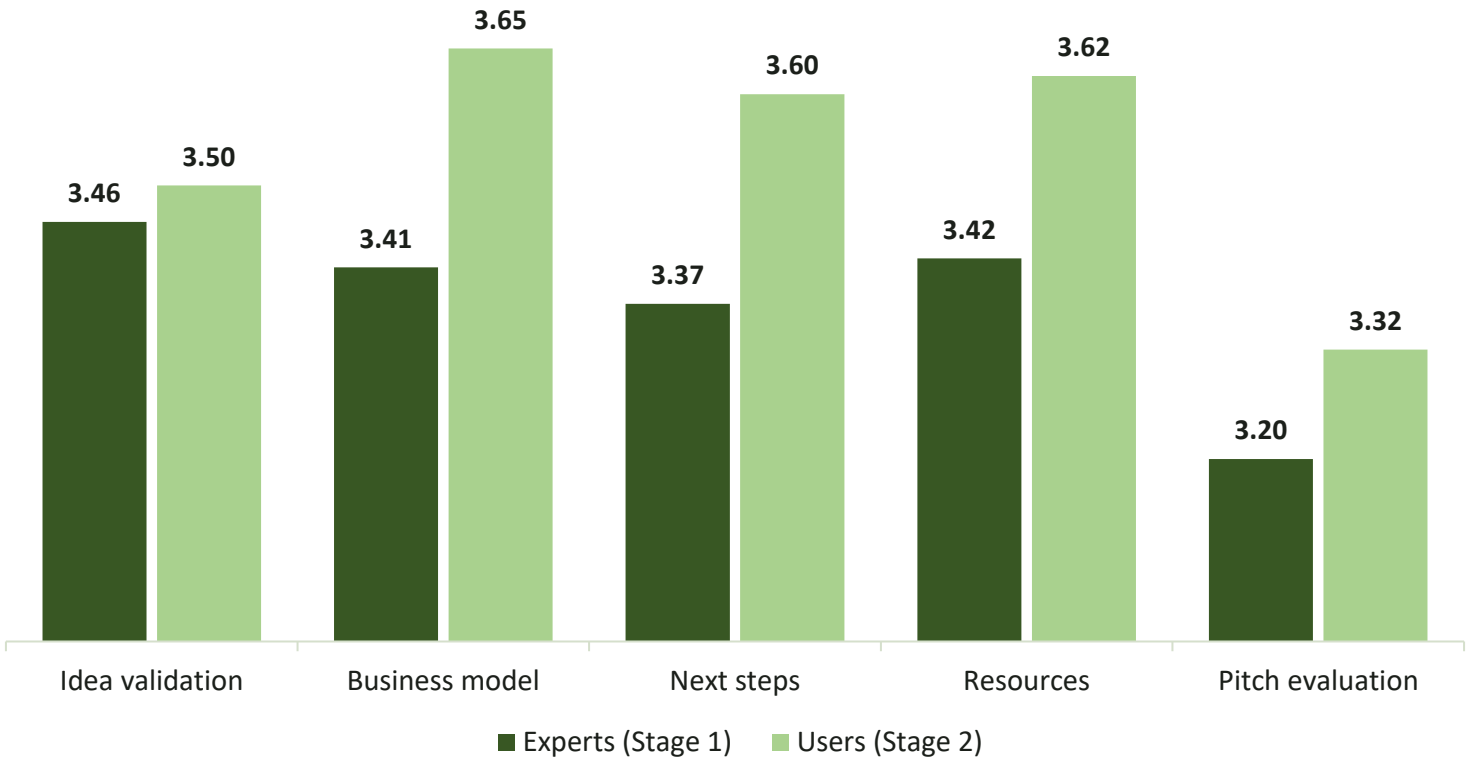
PAIRED COMPARISON · pitch vs mean of the other four areas

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

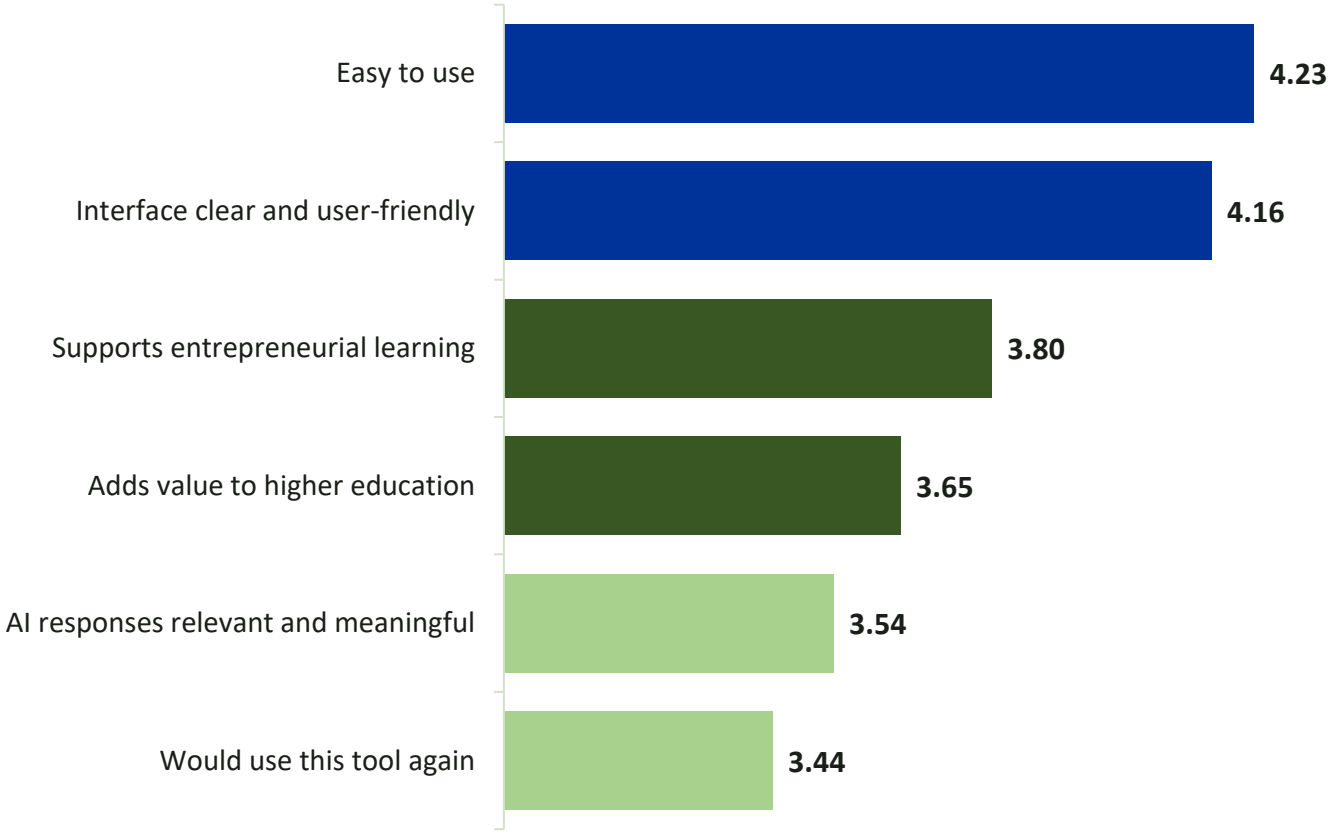
The two stages differ in

- 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.

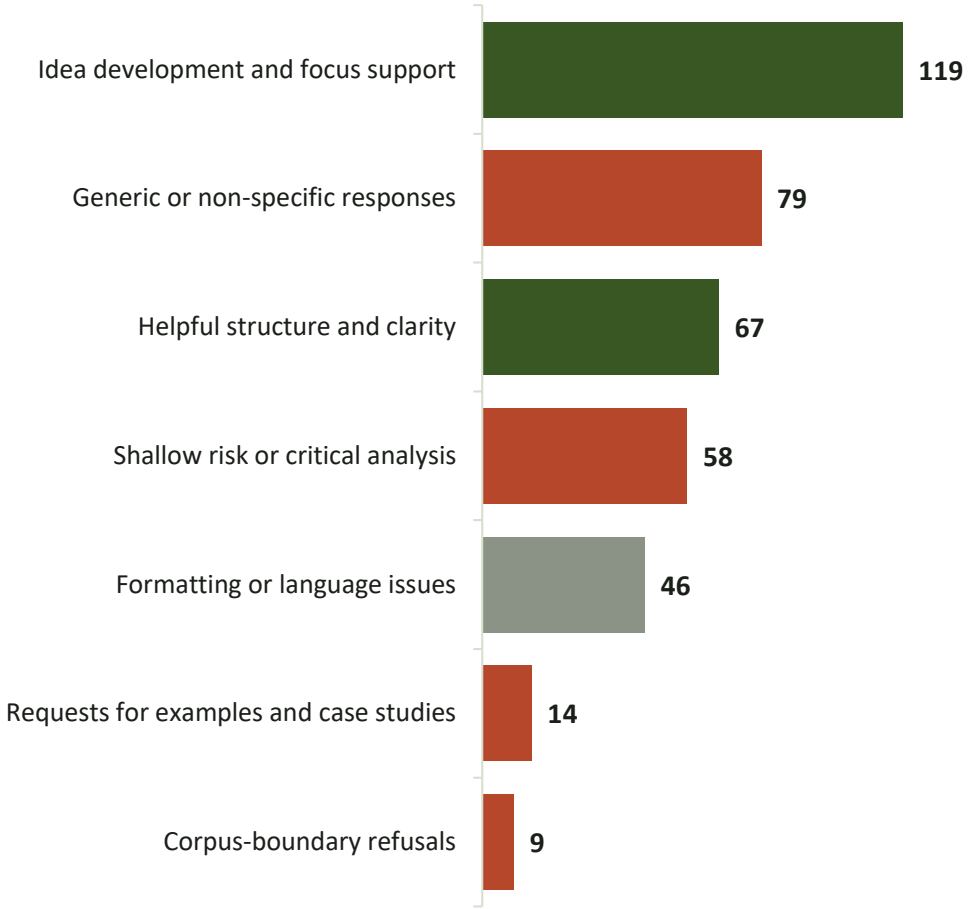
ADVICE

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

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

CONCLUSIONS

Evaluate functions, not averages

1. **Adequate content, high usability, one weak function**
   Expert mean 3.37/5; ease of use 4.23; pitch evaluation lowest in both stages, $t(113) = -3.19$, $p = 0.002$.

2. **The pitch gap is invisible in aggregate**
   It appears only with disaggregated, criterion-level measurement from two independent sources.

3. **Usability is not the bottleneck**
   Reuse intention tracks content quality, not interface ratings; improvement effort belongs in the corpus.

4. **Corpus composition is a pedagogical decision**
   Knowledge that lives in practice — worked pitch exemplars — must be deliberately represented; the proposed remedy is low-cost and measurable with the existing instruments.

Limits: affiliated (not blinded) expert raters · convenience field sample, no generic-chatbot control · perceived quality, not learning outcomes · one deployment, one corpus, one model.

Evaluation instruments (expert & field)
github.com/astmarinov/SEEN-Surveys · platform: seen.tu-varna.bg

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

Co-funded by the European Union (Erasmus+, SEEN). Views are the authors' own.

Co-funded by the European Union