

Experimentation and AI grading assistance

Blue Train

15th October 2026

Change.



Your way.

1

Context of the experimentation



Context of the experimentation

- **Observation** : exam grading considered as a major pain point for faculty members
- **Audencia** : a pioneer in exploring and testing AI assisted grading solutions since 2023
- **Pilot project objective** : assess the potential time savings in exam grading and the quality of feedback provided to students
- **Choice for the POC** : start up Blue Train, based on high standard quality, specialised on business schools and labelled Deeptech since Spring 2026

Project scope

Sampling : multiformat, multiprograms, FR/EN

- **Free POC fall 2025 : premaster Fondamentaux du Droit 35 copies**
- **POC 5 crédits Spring 2026**

Program	Cohort for testing	Topic	Exam type
MSc	50	Data for marketing	Essay with Python
MSc	70	Valuation techniques	Complex essay with analysis tables, text, excel
MS	45	Web analytics	Powerpoint with images, graphs
Bachelor 3A	20	Sales	Negotiation oral

**+ 1 credit « master thesis » on marketing / conseil / finance
on literature review deliverable**

2

Professor feedback



Jose Romero MORENO – Python exam and literature review

Eric Casenave – Sales oral exam

Bluetrain test · Negotiation and sales oral exams · Bachelor

An AI second grader for our sales oral exams

Today

45 students, a 25-minute role-play, 3 examiners.

5 minutes to grade and 3 lines of comments at most.

Students get a few lines of feedback on 2 or 3 criteria.

What we tested

- 1 An examiner **plays the buyer**, grades as a jury member and records the interview on a smartphone (GDPR consent)
- 2 **Send** the audio file to Bluetrain
- 3 The AI **transcribes, grades against our grid** and backs each criterion with quotes from the interview
- 4 The professor **compares and sets** the final grade

OUR GRADING GRID · 9 CRITERIA, OUT OF 20

Materials /1 (jury only)

Contact /1

Discovery /2

Argumentation /2

Objection handling /3

Negotiation /4

Closing and agreement /1

Assertiveness /3

Self-debrief /3

Protocol: 3 interviews picked by grade (17, 18, 8), 16 at random, 19 in total



Eric Casenave (2026)

First test · same interview, 17/20 from both the jury and the AI

Same grade, two levels of feedback

JURY · WRITTEN IN 5 MINUTES

« Très bonne performance, notamment sur la partie négø. Elle a vraiment obtenu une situation win-win, grâce à une compréhension de mes enjeux et sa force de conviction. Attention à des mots parasites comme “genre” qui détonnent avec son attitude pro. »

~40 words · 1 strength, 1 point of attention

AI · EXCERPTS FROM ~350 WORDS OF FEEDBACK

Opening	« Dès l'ouverture, vous installez la confiance », backed by a quote from the student
Strength	« Face à l'objection du concurrent à 6 000 €, vous avez le bon réflexe de creuser avant de répondre »
To improve	« Vous dégainez immédiatement vos références [...] sans d'abord explorer la crainte réelle du DRH »
Closing	« C'est votre interlocuteur qui a récapitulé les termes de l'accord »
Levers	Ask « Qu'est-ce qui vous ferait vous sentir en confiance sur ce point ? », then summarise the deal yourself

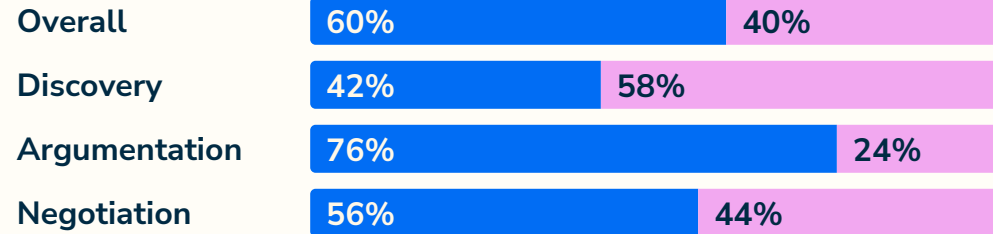
Complementary: the jury captures the overall impression and filler words; the AI backs each criterion with quotes and gives concrete levers.

Version 2 - requested additions: buyer profile, talk time, filler words

What the jury cannot measure live

TALK TIME BY PHASE

■ Seller (student) ■ Buyer (jury)



The student talks 76% of the time in argumentation, consistent with the AI's diagnosis: descriptive arguments, no proof.

Buyer profile identified

Novelty (scenario S2)

Filler words counted

« un petit peu » more than 12 times, « génial, génial, génial », « du coup », « en fait »

Missed signal

The buyer's need for novelty, « que vous ne captez pas suffisamment »

JURY, SAME INTERVIEW

« Très bon départ : contact, questionnement, écoute, validation (synthèse absente). Des difficultés lors de l'argumentation (remise) qui ont terni la prestation. Faible résistance aux demandes de baisse. »

Results - 19 interviews graded by the jury and by the AI

AI grades converge with the jury's

0.87

correlation between jury and AI grades

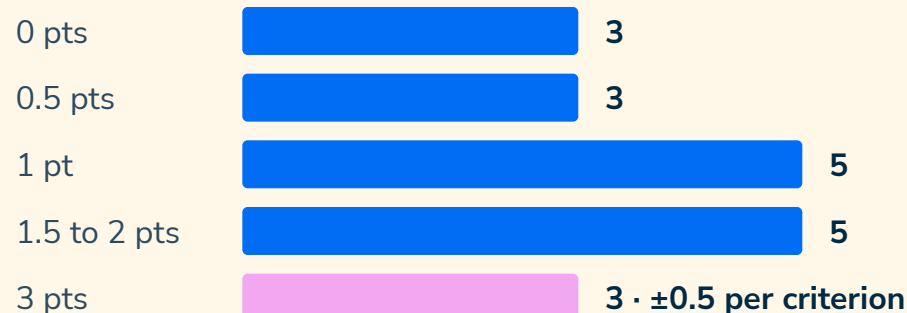
1.3 pts

mean absolute gap, out of 20

16 / 19

grades within 2 points of the jury's

Jury / AI gap, number of interviews



Same diagnosis, criterion by criterion: the 3-point gaps are half-points adding up. Example: 7/20 vs 10/20, both 1/4 for negotiation.

A benchmark to align graders: the AI applies the grid the same way to every student.

To be tested: extreme cases. Would a student who ignores the method and has nothing to say get 2/20?

Benefits, potential and limitations

Richer assessment, professors decide

For students

Feedback criterion by criterion, quoting their own words

Buyer profile, talk time, filler words

A grade students see as less prone to halo effect

For professors

The whole grid covered, even overlooked criteria

A benchmark to align graders

A proposed grade the professor validates

For the school

A cohort report: strengths and gaps of the class

Data for AACSB, etc who demand feedback

Models improve each year; starting now builds our expertise for future simulators

Known potential: ~430 students

All sales courses with an oral exam: Bachelor year 1 Paris and Nantes
~300 · Bachelor year 3 ~80 · MSc ~50 (to my knowledge)

Limitations

3 of 45 students refused to be recorded (RGPD)

We record by hand and send files to Bluetrain: no instant feedback

2

Perspectives



Overall results



Essays / exams

Overall time savings from **50 to 70% estimated**

Key success factors : a **precise criteria** grid set between the professor and Blue train team

All tested formats worked technically speaking : essays, fichier excel, pwts, images...



Orals

Gain : AI is used as an assistant and is a good complement to the jury on **objective** criteria, « boosting » the evaluation

= qualitative gain on objective criteria (not time savings)

Would need additional POCs to be explored



Master thesis results

Mitigated result, needing a bigger POC

AI can be efficient with specific criteria on **mechanical** aspects

So far, AI non efficient on the creative work, link with the human experience etc (luckily 😊)

Synthesis

Positive

- **Scalable tool (big cohorts) : different types of essays / "objective" criteria on evaluation grid**
- **Interface with LMS**
- **Human supervision and validation is the core**
- **Qualitative coaching of the Blue Train teams**
- **AI Act / RGPD compliance**

Areas of improvement

- **As per mini test on literature reviews, lack of efficiency on creative writing and thinking, state of art, etc (good news :) for humanity**
- **One professor mentioned he would have enjoyed to access the AI model directly to work it out**

Question : with AI fast development, data security is going to be at heart of any AI POC

THANK YOU 😊

