Typist, Not Interpreter: Measuring and Improving AI Accuracy in Transcript Data Extraction for Credential Evaluation

Veri Eval White Paper · November 2026

About this paper

Author: Temitope Petrus-Nihi, Senior Director, Credential Evaluation, Veri Eval

Tests: Microsoft AI Builder (GPT-4.1 mini, GPT-4.1, GPT-5 chat) on five real, redacted transcripts in October and November 2026; Microsoft AI Builder and the ChatGPT, Claude and Gemini chat apps on ten fictional transcripts, 20 to 23 November 2026, with three prompt versions and an answer key for every line.

Acknowledgments: The Veri Eval IT Team and the Veri Eval Research Team.

License: This paper and its free resources are released under Creative Commons Attribution 4.0 (CC BY 4.0). Anyone may share and adapt them, with credit to Veri Eval LLC.

Introduction

Foreign academic transcripts must be transcribed into a course-level record before evaluation can begin, a task that most offices continue to perform manually. Although AI tools are increasingly promoted as a means of automating this step, and have been demonstrated at the field's conferences, no published study has measured how accurately they copy real foreign transcripts or described how an office might detect their errors.

This paper reports the testing that Veri Eval, a foreign credential evaluation company based in The Woodlands, Texas, conducts on real casework before adopting any tool for its evaluation platform. AI models in Microsoft AI Builder and in the ChatGPT, Claude and Gemini chat applications were tested on five real and ten complete fictional transcripts in six languages, with each course line scored against a verified answer key.

The paper is intended for foreign credential evaluators and for the admissions directors and evaluation service managers who determine whether and how AI is used in their offices. It reports the findings, describes the principal types of error, and provides free

tools with which any office may test and check its own AI.

Summary

The highest accuracy was obtained with a short "typist, not interpreter" prompt used with an AI tool that had been tested on the type of transcript it would process. With this prompt, the ChatGPT, Claude and Gemini chat applications copied between 98.2% and 99.6% of course lines exactly across ten fictional transcripts. Microsoft AI Builder (GPT-4.1 mini), the tool used by Veri Eval, copied 96.1% of the same lines and 75.3% of the lines on five real, redacted transcripts from Liberia, Kenya, Pakistan, Bangladesh and Afghanistan.

Larger models did not perform better. Within Microsoft AI Builder, GPT-4.1 and GPT-5 chat copied only about 61% of the fictional lines and 36% to 39% of the real ones, well below the smaller GPT-4.1 mini. Of greater concern for evaluators, when these models misread a grade, the misread grade was higher than the printed one in about four cases out of five, an error that would raise an applicant's GPA without any visible sign.

The rules in the prompt affected accuracy in both directions. When the tools were given no rules, only the output format, they read the transcripts correctly but altered what they had read, for example merging a mark and a letter into "C (51)" or reversing bilingual course titles. When further rules were added to the typist prompt, the tools began to interpret rather than copy: one left an entire credit column blank after concluding that the column did not contain credits. The sums the tools wrote into their notes were also unreliable: ChatGPT, Gemini and Microsoft AI Builder got most of theirs wrong, although Claude's were correct.

It follows that AI tools should be used only to copy, never to calculate. Simple spreadsheet checks against the totals and GPAs printed on the transcript itself detect most of the errors that remain, but not all of them: when two grades are exchanged between rows, every total is unaffected, and each line must therefore still be read by the evaluator. Taken together, these recommendations form what this paper calls the TYPE method (Typist prompt, Your own test first, Printed-totals check, Evaluator decides).

Checking an AI-filled record was substantially faster than typing one. A staff member required 26 minutes 42 seconds to type a 52-line transcript by hand, whereas checking AI-filled records of 64 and 84 lines required 9 to 13 minutes each, and every incorrect course line was found and corrected. An AI tool therefore need not be perfect, provided that the check is thorough; the office's requirement is a careful checker rather than a flawless tool.

Warning an AI tool against a particular mistake can induce that mistake. Of the four rules added to the prompt, two improved the results and two worsened them: a rule describing the points column led one model to copy points as grades on 12 lines, and a rule asking the model to total the credits produced an incorrect sum on every occasion. Arithmetic belongs in the checks, not in the tool.

To enable any office to test its own AI tools, the paper provides the prompts tested, a free spreadsheet checker, and ten free fictional test transcripts in six languages, each with its answer key.

What this means for your office

Together, these recommendations make up what this paper calls the **TYPE method**: a **T**ypist prompt, **Y**our own test first, a **P**rinted-totals check, and an **E**valuator who reads every line and decides.

1. **Use AI as a typist, not an evaluator.** An AI tool may draft the course record, but it should not judge equivalency, convert grades or make any other decision.
2. **Use a short typist prompt.** The prompt should instruct the tool to copy only what is printed, exactly as printed, never to convert or calculate, and to leave a field blank and flag it when anything is unclear. The prompt used in this study may be copied freely (Appendix A). With no rules, AI tools tend to rewrite; with too many, they tend to interpret.
3. **Do not ask AI tools to perform arithmetic.** The tools' own sums were unreliable: ChatGPT, Gemini and Microsoft AI Builder got most of theirs wrong, even where the lines behind them were correct, while Claude's were correct. Because an office cannot tell in advance which kind of tool it has, arithmetic should be performed in a spreadsheet.
4. **Select an AI tool by testing it, not by its size or name.** The ten free fictional transcripts can be processed by the tool and the replies scored against the answer keys. The same prompt achieved 61% in one configuration and 99.6% in another. The tool should be retested after any change to the prompt, the tool or the model.
5. **Check every record before use.** The extracted lines should be compared with the credit totals, points and GPAs printed on the transcript itself. The free spreadsheet performs this comparison (*Checks that catch the errors*).
6. **Read every row as well.** Totals do not reveal every error: two grades exchanged between rows leave every total and the GPA unchanged. The list in *What to check for* is intended for this step.

7. **Let a system, not a person, transfer the files.** Copying files and replies between windows by hand invites errors such as a reply saved under the wrong name. A system that submits the transcript and saves the reply itself avoids this source of error.
8. **Protect applicant records.** Only business AI accounts whose contracts protect the office's data should be used, never personal or free accounts (*Protect applicant data*).

Terms used in this paper

Term	Meaning
Prompt	The written instructions submitted with each transcript
v1.1, v1.2, v1.2b	The three versions of our prompt: the original, one with four added rules, and one with only the two rules that helped
Model	The underlying AI engine, such as GPT-4.1 mini; the same tool can offer several
AI Builder	Microsoft's tool for using AI inside Power Platform, the system our office runs on
Chat app	ChatGPT, Claude or Gemini used through their ordinary chat window
Temperature 0	A setting that makes the tool's answers as consistent as possible from run to run
JSON	A structured text format that a spreadsheet or system can read; the tool's reply comes back in it
Run	One transcript given once to one AI with one prompt
Course line	One course on the transcript: its code, title, credits and grade
Right line	A course line where the code, title, credits and grade all match the answer key exactly
Answer key	The correct record of a transcript, checked against every figure the transcript prints

Term	Meaning
Printed totals	The credit totals, averages and GPAs the transcript itself prints, used to check the tool's lines
Header fields	The details above the course list: name, institution, program, dates, grading key
Predictions fixed in advance	What we expected to find, written down and dated before any test, so the results could not be fitted afterward

Why this matters

In credential evaluation, transcription precedes every other step. Grades cannot be converted, nor averages calculated, until each course line (its code, title, credits and grade) has been entered into the system, and on a long transcript this entry is time-consuming and prone to error. Vendors, conference sessions and webinars now advise offices that AI can perform this transcription, and many offices have begun to adopt it.

Evidence relevant to this kind of work, however, is scarce. Published accuracy studies concern US transcripts in English, or open-source models tested on applicant records that cannot be shared, while vendors report accuracy figures without describing how they were obtained.

To the author's knowledge, no study from within the credential evaluation profession has measured the accuracy of AI extraction from scanned foreign transcripts with varied layouts, using the commercial tools that offices actually use. Nor has any study classified the types of error involved, proposed prompt revisions and checks, compared the time required with that of manual typing, or provided a shareable test set.

The present paper addresses each of these gaps. It is presented as a method that any office can apply with any AI tool, rather than as a description of a single company's system.

What's already known

Measurements of AI transcript extraction have been published, but, as the following table shows, none corresponds directly to the work of a foreign credential evaluation office.

Source	Transcripts	AI tested	Reported result
Schubert & Bießmann, 2026	100 from international applicants to a German master's program	Open-source models only	Best F1 about 0.51; grades harder than credits; data not shareable
Torkian & Zhou, 2026	40 US high school, English only	Multi-agent system	96.7% against expert review; international formats excluded
ACM Learning @ Scale, 2025	US transfer admissions	AWS Textract, GPT-4o, Claude 3.7	About 90% for the best pipeline
Lowry, LSAC	US undergraduate	Text recognition (OCR)	About 98% per character; experts found typing grades quicker than fixing errors
RTI International	US college	Commercial extraction + human check	Entry time 10.7 → 7.5 minutes per transcript

The German study is the closest to the present work: it concerns international applicants, it found grades more difficult to extract than credits, and its data could not be shared because they are personal records. The fictional test set described here is designed to overcome this last limitation.

Vendors publish figures without describing their methods, ranging from "90%+" accuracy in a conference demonstration (Tondro Consulting) and "up to 99%" (DegreeSight) to the claim that AI extraction "virtually eliminates errors" (Verionce).

The profession's own guidance sets out principles rather than measurements. TAICEP's position paper (2025) establishes ethical principles, AICE (2024) asks its members to test their AI platforms regularly, and a Council of Europe survey (2026) of 138 institutions found AI use to be limited and closely supervised. The free test set is intended to give offices a practical means of applying these principles.

Checking extracted data against printed totals is not in itself new: one vendor demonstration describes calculation-consistency checks, and the 2026 US study uses the GPA as an indicator of reliability. The contribution of this paper is to make such checks

open, tested and freely available to any office.

Method

Throughout the study, the tools were instructed only to copy transcripts, never to interpret them, and every course line was scored against an answer key verified by the transcript's own printed totals.

The real transcripts

The study began with five real applicant transcripts, redacted and selected for their differing layouts. All five were scanned documents, and they are described here by country and layout only (see *Privacy*).

Transcript	Layout	Course lines	Credits
Liberia	Two course columns; one page of a longer record; quality points printed	21	56 (page prints 137 for the whole record)
Kenya	One column over four pages; academic hours, no credit total printed	65	3,466 hours
Pakistan	Two columns side by side	52	135
Bangladesh	Four year-columns; term totals and term GPAs printed	73	161.5
Afghanistan	Mark column beside a points column (credits × mark)	84	204

The fictional transcripts (test set v2.0)

Ten complete fictional transcripts were then constructed from invented institutions. Each is marked "SPECIMEN — FICTIONAL" on every page and carries no seals, crests or signatures. Five reproduce the real layouts as complete records: two columns (C1), a single column with hours and no credit total (C2), term blocks arranged two across (C3), four year-columns with a supplementary term in small print (C4), and a mark column beside a points column (C5). The remaining five introduce further languages, scripts and

grading scales: Spanish on a 0–10 scale with comma decimals (R1), French marks out of 20 with coefficients (R2), Chinese and English with fractional credits such as 2.5 (R3), Arabic and English in a right-to-left layout (R4), and German on the inverted 1.0–5.0 scale (R5). Together they contain 561 course lines.

Each fictional transcript exists in a scanned-appearance version, which was used for testing, and a clean version, accompanied by an answer key in JSON and a readable form. A separate script confirmed that every line in each key is printed on the page and that every printed total, average and GPA (228 figures in all) can be recomputed from the key lines. Where a document prints a value in two forms, such as a mark and a letter or a title in two languages, an exact copy of either form is scored as correct. The five short single-error specimens from the first draft are retained as teaching samples (Appendix B) and are included in the free test kit, together with the altered copy of Specimen 03, but are not scored.

The prompt

The instruction text (a structured prompt) asks the tool to return each transcript in a fixed set of fields. To determine how much instruction is required, the prompt was tested at three levels, ranging from no rules to an extended set:

Version	What it adds	Where tested
v0, format only	The output fields and "return only the JSON"; no rules at all	Chat apps, ten fictional transcripts
v1.1, typist	Ten copying rules: copy only what is printed, exactly as printed; never convert, never calculate; leave a blank and flag what is unclear	Everywhere; Veri Eval's live prompt
v1.2b, extra rules	v1.1 plus two rules written from failures: no calculated numbers anywhere, and read side-by-side blocks one at a time, checking each block's printed total	Everywhere on the fictional set

An earlier version, **v1.2**, added four rules to v1.1 and was tested on the real transcripts and on five of the fictional ones. The full text of every version is given in Appendix A.

Tools and models

- **Microsoft AI Builder** (the prompt feature of Microsoft Power Platform, and the tool used in Veri Eval's evaluation system), run in a separate test environment at temperature 0, the setting that makes responses as consistent as possible. Models: GPT-4.1 mini, GPT-4.1 and GPT-5 chat. Real transcripts: all three models with v1.1 (by 29 October 2026) and GPT-4.1 mini with v1.2 (by 2 November 2026). Fictional transcripts, 20 November 2026: GPT-4.1 mini with v1.1 and v1.2b (all ten) and with v1.2 (five); GPT-4.1 and GPT-5 chat with v1.1 (five).
- **ChatGPT** (free plan, default model; no model name was displayed), **Gemini** (free plan, "Auto" setting) and **Claude Sonnet 5.5** (Medium effort, the model offered on the free plan), each tested on all ten fictional transcripts with v0, v1.1 and v1.2b between 20 and 23 November 2026. Every run used a new private or temporary chat with memory disabled, and the prompt and PDF were submitted in a single message.
- One run was made for each combination of transcript, model and prompt, giving approximately 125 runs on the fictional set.

Model names and application settings are reported as they appeared on the dates of the runs. Because these tools change frequently, the results describe those versions rather than the products in general. Some replies copied from the chat applications contained formatting characters added by the applications' copy buttons, such as a backslash at the end of each line; these were removed by a fixed cleaning step before scoring, and the original replies were retained. Course titles were compared after standard character normalization, so that a full-width bracket and an ordinary bracket are treated as the same character. All scores are reported under this strict rule. Under a more lenient rule for bilingual titles, which accepts a title containing every language part of the printed title in any order, only three results change: Claude with no rules rises from 402 to 452 lines, and Gemini gains two lines under v1.1 and v1.2b.

Pass marks, set before testing

1. At least 95% of course lines copied exactly.
2. The credit total matched on at least 90% of transcripts.
3. No invented courses.
4. Less time required than typing by hand.

Predictions fixed in advance. Before any run on the fictional set, 22 predictions about the results were recorded (concerning which layouts would prove most difficult, which prompt would perform best, and which errors the checks would detect) and fixed on 18

November 2026. The dated document was sent by email on that day, so that its timestamp establishes its precedence. Each prediction is reported as held or not held, including those that failed, and none was altered after the first run.

Answer keys and scoring

For the real transcripts, the key was read line by line and verified against every figure printed on each transcript, including term credit totals and GPAs, semester point totals and overall totals. Thirty-two lines could not be verified against any printed figure (see *Limits*).

A line is scored as correct only when its code, title, credits and grade all match the key, and lines are matched to the key by code, title and order. The header fields (name, institution, program, dates, grading key and printed totals) are scored separately, each as exact, correct apart from capitalization or spacing, or incorrect.

Time study

Each AI run was timed from submission to the complete reply, which in every case arrived within one minute; the time study therefore concentrates on the human component of the work.

Typing by hand was estimated for three transcripts (C2, C3 and C5) and measured for one of them. The estimate counts every keystroke required to type the complete record from the answer key, including the header fields and every course line (approximately 4,400, 2,700 and 4,400 keystrokes), and divides this count by published typing speeds: an average of 52 words per minute, or approximately 15,600 keystrokes per hour, from a study of 168,000 typists (Dhakal et al., 2018), and a data-entry industry average of approximately 11,100 keystrokes per hour. Because typing tests do not account for the time spent reading a scanned transcript, keeping one's place and re-checking, a staff member who had not previously seen the transcripts also typed C3 by hand in a single timed sitting. The ratio between her measured time and the C3 estimate represents this additional time, and the same ratio is applied to C2 and C5.

Checking AI output was measured directly. After typing C3, the same staff member checked and corrected four AI-filled records: the most and least accurate Microsoft AI Builder replies for C2 and for C5. She was not informed which tool had produced each record and had no access to the answer keys. She checked one record of each transcript on 20 November 2026 and the other on 23 November 2026, in reverse order, so that each transcript was checked once with errors and once without and familiarity could not favor

one condition. C3 was not checked, as she had already typed it. The measure is the time required to produce a correct record, together with any errors remaining after checking. No run times were recorded for the initial test on the real transcripts.

Privacy

The real transcripts are not reproduced anywhere in this paper. They are described only by country and layout, no names, numbers or institutional details are quoted, and all illustrations use the fictional specimens. The applicants did not consent to publication, redaction can leak information, and legible images of real transcripts could assist forgers.

Results

No model met the pass marks on the real transcripts, and the smallest model, GPT-4.1 mini, was the most accurate by a wide margin.

Real transcripts: course lines copied exactly (prompt v1.1)

Transcript	GPT-4.1 mini	GPT-4.1	GPT-5 chat
Liberia, two columns	21/21 (100%)	17/21 (81%)	19/21 (90%)
Kenya, one column	65/65 (100%)	59/65 (91%)	58/65 (89%)
Pakistan, side by side	35/52 (67%)	13/52 (25%)	15/52 (29%)
Bangladesh, four year-columns	35/73 (48%)	15/73 (21%)	3/73 (4%)
Afghanistan, mark + points	66/84 (79%)	11/84 (13%)	11/84 (13%)
All five	**222/295 (75.3%)**	**115/295 (39.0%)**	**106/295 (35.9%)**
Credit total matches	2 of 5	1 of 5	2 of 5
Invented courses	0	0	0

Grade errors exceeded credit errors for every model (Appendix C1). For GPT-4.1 mini, the credit total proved a reliable indicator on these transcripts: every transcript whose extracted credits summed to the printed total was entirely correct, and every transcript whose credits did not contained errors.

These figures are more stringent than most published accuracy rates, because a line is counted as correct only if all four of its values are exact, and because the transcripts were scanned documents with multi-block layouts.

Improved prompt (v1.2) on the real transcripts

GPT-4.1 mini was then run again on the five real transcripts with prompt v1.2. It copied **215 of 295** lines correctly (72.9%), compared with **222** (75.3%) under v1.1; the revised prompt thus performed slightly worse overall (results by transcript in Appendix C2).

Under v1.2, missed lines fell from 8 to none and the credit total matched on three transcripts rather than two, but grade errors rose from 58 to 75 and the number of lines flagged as uncertain rose from 1 to 14. Neither version invented a course, although v1.2 listed one course twice. The effect of each rule is set out under *Did the rules work?* below.

The fictional transcripts

Across the ten fictional transcripts, the short typist prompt (v1.1) was the best prompt, or within two lines of the best, for every AI tool tested, and it was the only prompt that never caused a large drop in accuracy. With this prompt, the three chat applications copied between 98.2% and 99.6% of lines exactly, and Microsoft AI Builder (GPT-4.1 mini) copied 96.1%, as shown in the table below. Gemini was the exception to the general pattern: its three prompts finished within two lines of one another (551 to 553 of 561), so for that tool the prompt made little difference.

AI (all lines scored, 561 per full set)	v0, format only	v1.1, typist	v1.2b, extra rules
Claude Sonnet 5.5	402 (71.7%)	**559 (99.6%)**	559 (99.6%)
ChatGPT (free)	370 (66.0%)	**558 (99.5%)**	494 (88.1%)
Gemini (Auto)	553 (98.6%)	551 (98.2%)	553 (98.6%)
Microsoft AI Builder (GPT-4.1 mini)	not tested	**539 (96.1%)**	531 (94.7%)
Microsoft AI Builder (GPT-4.1), five transcripts	not tested	210 of 345 (60.9%)	not tested
Microsoft AI Builder (GPT-5 chat), five transcripts	not tested	212 of 345 (61.4%)	not tested

Microsoft AI Builder (GPT-4.1 mini) was also run with v1.2 on five transcripts: 319 of 345 (92.5%).

No AI tool invented a course in any run, consistent with the results on the real transcripts. Where lines were incorrect, the grade was usually the cause, and the types of grade error differed considerably between tools:

AI on v1.1	Lines with the wrong grade	What happened
ChatGPT	0	
Claude Sonnet 5.5	2	Both "bestanden" (passed) left blank, not guessed
Gemini	4	Two grades swapped between neighboring rows; two "bestanden" left blank
Microsoft AI Builder (GPT-4.1 mini)	19	Mostly a dropped plus or minus (A- read as A)
Microsoft AI Builder (GPT-4.1), five transcripts	117	Mostly read higher than printed
Microsoft AI Builder (GPT-5 chat), five transcripts	111	Mostly read higher than printed

Some of these runs were not included in the plan fixed on 18 November. When the first results indicated that the prompt affected different tools differently, the v0 and v1.1 runs in the chat applications, and the ChatGPT and Claude runs with v1.2b, were added so that each application could be compared with itself.

Header fields

On the real transcripts, only the printed totals could be scored, as no key was prepared for the other header fields. Both prompt versions copied the printed credit total correctly on all five transcripts, four exactly and one in a different format. For the printed average or CGPA, v1.1 was correct on four of five: on the Kenyan transcript it copied 64.00, the cumulative average printed on earlier pages, instead of the final figure of 65.00. v1.2 was correct on all five.

On the fictional set, header details (name, institution, program, dates, printed totals, language and grading key) were scored across all 125 replies. Of the 1,125 fields, 85.0% matched the answer key under the strict rule and 94.2% under the lenient rule, against 89.0% for course lines. The main weakness was on the bilingual transcripts, where most tools dropped the non-English half of the name, institution and program. The full breakdown is in Appendix C3.

Time

Checking an AI-filled record required roughly one fifth to one third of the time needed to type the same record by hand, and the checker found and corrected every incorrect course line the tool had made.

Task	Transcript	Lines	Time	Result
Typing by hand	C3	52	26 min 42 s	Time only
Checking an AI record with no errors	C2	64	9 min 57 s	No line errors added; final mean mark replaced by a year mean
Checking an AI record with 11 wrong lines	C2	64	13 min 20 s	All 11 found and fixed
Checking an AI record with 18 wrong lines	C5	84	10 min 42 s	All 18 found and fixed
Checking an AI record with no errors	C5	84	9 min 8 s	No errors added

At the measured typing speed, the 64-line C2 and the 84-line C5 would each require approximately 44 minutes to type, against 9 to 13 minutes to check. Most of the checking time is spent reading every cell, so a clean record still took about 9 to 10 minutes; records with errors took longer on both transcripts, by 3 minutes 23 seconds on C2 and 1 minute 34 seconds on C5. In practice, a checker therefore cannot tell a good record from a poor one at a glance, and every record needs the full check. Every AI reply, by comparison, arrived within one minute.

Predictions

The 22 predictions were fixed and dated on 18 November 2026, before the final runs, and are marked in Appendix C4 against the strict scores. Seven held, three held in part and twelve did not hold. That most did not hold is itself informative: the predictions recorded what the authors expected from earlier work, and the results departed from those expectations in ways that matter for practice.

Three results deserve comment. The false alarms (P13) arose on C2, where six replies copied the letter grade rather than the mark, so no mean could be worked, and on R1, where one reply attached grade words to the numbers; both are cases where a check fails for want of a number rather than because a number is wrong. The grade errors the checks missed (P12) were chiefly pass-only modules left blank on R5, which carry no weight in any average, and letter grades added beside correct marks on R4. Finally, wrong letter grades on C1, C3 and C4 were higher than the printed grade in 193 of 236 cases. Had the tools substituted another grade from the same transcript at random, about 130 would have been higher. The tilt appeared on all three transcripts, though it came mainly from the AI Builder models, which made most of the grade errors. Errors of this kind favor the applicant, which strengthens the case for checking every reply.

How AI gets it wrong

The principal risk lies not in invented courses but in plausible incorrect values, which appear correct unless compared directly with the transcript.

Error type	What happened	Where it struck	Why it matters
Grade wrong by a modifier	A− read as A, A+ as A, B as B+	GPT-4.1 mini: 45 grade errors against 1 credit error on the side-by-side and year-column layouts	Changes the GPA; invisible to a credit total
Whole term skipped	Eight courses of one term left out	GPT-4.1 mini, four year-columns	Credit total falls short
Stopped early, without saying so	19 of 73 lines returned	GPT-5 chat, four year-columns	Looks like a complete record

Error type	What happened	Where it struck	Why it matters
Points copied as grades	The points column (credits × mark) put in the grade field on every line	GPT-5 chat, mark + points layout	Every grade on the transcript wrong
Repeated values	The same grade or number copied down long runs of lines	GPT-4.1: A+ down a column; "32" or "16" on 26 lines	Grades the transcript does not show
Wrong mark or credit	A mark or credit value differing from the printed one	GPT-4.1 mini: 18 of 84 lines, mark + points layout	Plausible numbers; needs the source to catch
A printed figure miscopied	An average printed as 65.00 copied as 64.00	GPT-4.1 mini, one-column layout	Even pure copying slips

Errors were concentrated on the more complex layouts, which is also where the models differed most. On the two simple layouts, GPT-4.1 mini copied every line correctly; where courses are arranged in blocks side by side, or where a points column appears beside the grade, accuracy declined for every model.

The models' own warnings provided no safeguard. Across 15 runs, 442 lines were incorrect or missing, yet only 2 were flagged as uncertain, and no run detected that its credits did not sum to the printed total.

Larger models were not more faithful copyists: the smallest model made fewer errors than either of the larger ones. On the real transcripts this finding rested on one prompt, one run per model and one tool, and it was therefore treated as a hypothesis for further testing; the fictional set, described below, confirmed it.

The fictional set confirmed these patterns across four different AI tools and revealed four further types of error:

- **Grades read too high.** When the larger models in Microsoft AI Builder misread a letter grade, the misread grade was higher than the printed one in about four cases out of five (78 of 96 for GPT-4.1 and 75 of 91 for GPT-5 chat), so that a C+ could become an A. Because such errors favor the applicant, they inflate the GPA without any visible sign.
- **Grades exchanged between rows.** Gemini assigned two adjacent courses each other's grades. Because both grades remained present, every credit total and the GPA were unaffected; no arithmetic check can therefore detect this error, and only

reading the rows can.

- **The tools' own sums are unreliable.** The tools sometimes added sums to their notes, for example "courses add to 181; printed total is 179." ChatGPT got both of its sums wrong, Gemini 7 of 10 and Microsoft AI Builder (GPT-4.1 mini) 5 of 6, usually even though the lines they had copied added up correctly; in one note Gemini changed its answer in mid-sentence and settled on the wrong one. Claude's sums, about a dozen, were all correct. Because the reliability of a tool's arithmetic cannot be known in advance, the arithmetic belongs in a spreadsheet.
- **Blanks rather than guesses.** Every AI tool left the German pass mark "bestanden" blank rather than inventing a grade. This is the safer failure, because a blank field is readily noticed by a checker, whereas a plausible guess is not.

Prompt rules that fix them

For version 1.2, four rules were added to the prompt, each addressing an observed failure, while the structure of the reply was left unchanged. Appendix A gives version 1.1 in full, together with the exact wording of each change.

Rule	What it tells the tool	Failure it targets
Grade column	The grade is the mark or letter printed for the course, never a points or quality-points column (a number equal to credits × mark). Unsure which column is the grade → flag it uncertain and say why.	Points copied as grades
Repeated totals	If a total or average is printed more than once, take the final cumulative figure and say which page it came from.	Page and term totals mistaken for the whole record
Blocks side by side	Read one block completely, top to bottom, before the next; left to right, then down. Take credits and grade from the same row as the course name. Never skip a block. Where a block prints its own total, check it.	Skipped terms, grades taken from the neighboring block

Rule	What it tells the tool	Failure it targets
Final credit check	Add up the credits. If the sum differs from the printed total, write one note giving both numbers. Never change a course to make the numbers agree.	No run noticed its own mismatch

The last of these rules was intended to resolve a tension in the original prompt, which instructed the tool never to calculate. It permitted a sum in one location only, the notes field, where it could not alter any value.

Did the rules work?

Two of the rules improved the results and two did not. Overall, on the real transcripts, GPT-4.1 mini with v1.2 copied 215 of 295 lines correctly, compared with 222 under v1.1.

Rule	Effect	Evidence
Blocks side by side	Helped	Bangladesh's skipped term (8 lines) came back and its credits now add to the printed 161.5; Pakistan gained a line
Repeated totals	Helped slightly	Kenya's final average copied as 65.00, not the 64.00 printed on earlier pages
Grade column	Hurt	On Afghanistan, lines right fell from 66 to 56; on 12 lines the model put the points value in the grade field, the very mistake the rule names
Final credit check	Unreliable	The model's own sum was wrong or missing on all four transcripts that needed it, and once reported a match that did not exist

Two lessons follow for the design of prompts of this kind. **Naming a mistake can induce it.** The detailed description of the points column appears to have drawn the model toward that column, whereas rules stating what to do (take the grade from the same row as the course) proved more effective than rules describing what to avoid. **AI tools should not perform arithmetic.** The model's sums were incorrect on every occasion, while the same check performed by a spreadsheet or an intake workflow is exact; for this reason, the checks described in this paper are performed outside the tool.

On this evidence, the two effective rules were retained, the other two removed, and the result designated **v1.2b**; the production prompt remained on v1.1 in the meantime. Because a single run per transcript cannot establish whether small gains are genuine or due to chance, and because the rules were derived from the same five documents, v1.2b was subsequently tested on the fictional set.

On the fictional set, the question became how much instruction an AI tool actually requires. The answer proved to be neither no rules nor additional rules, but a short set of plain copying rules:

Prompt	What the tools did	Example
v0, format only	Read correctly, then rewrote what they read to their own taste	ChatGPT and Claude merged a mark and a letter into "C (51)" on all 64 lines of C2; Claude wrote "A (97)" on R4 and turned every Chinese-English title around
v1.1, typist	Copied what was printed	Claude 559 of 561, ChatGPT 558 of 561
v1.2b, extra rules	Began to interpret, audit and calculate	ChatGPT decided the C2 "Hours" column was not credits and left all 64 blank; Gemini said codes were printed back to front when they were not; Claude wrote ten notes checking the transcript's arithmetic, all correct; the sums ChatGPT and Gemini offered were mostly wrong

The copying rules, in other words, do not improve the tool's reading. Without them, every grade and credit that Claude chose to record was still correct; the function of the rules is to prevent the tool from rewriting what it has read. Rules added beyond them, by contrast, invited precisely the interpretation the prompt is designed to prevent, while correcting little that v1.1 had not already corrected. Microsoft AI Builder (GPT-4.1 mini) behaved in the same way: v1.2b corrected a few grades, but it also omitted, without notice, two blocks of courses that v1.1 had copied without difficulty.

The practical implication for an office is to keep the prompt short and plain, to change only one rule at a time, and to retest on a fixed set of transcripts before any change is put into production. Errors that remain should be caught by checks rather than by additional

rules.

Checks that don't trust the tool

The principle of the checks is straightforward: the extracted values are summed and compared with the figures printed on the transcript itself. Where the two disagree, the discrepancy identifies the block that must be re-read.

Three checks

Check	Compares	Catches	Needs
Credit total	Sum of extracted credits against the printed total and each printed term total	Skipped lines and terms, credit misreads, stopping early	Nothing beyond the transcript
Points total	Credits × mark, summed, against printed points totals	Mark misreads, points copied as grades	A transcript that prints marks and points
GPA	GPA worked out from the extracted grades against the printed term and final GPAs	Grades wrong by a modifier	A grading key printed on the transcript that gives grade points

The GPA check uses only the grading key printed on the transcript itself (for example "A 4.0; A− 3.7; B+ 3.3"). Where the key gives only ranges or descriptions, or where no key is printed, the GPA check is omitted, since no external grading scale is applied and no conversion is performed.

A credit total cannot detect an incorrect grade, because grades do not affect the sum of the credits. The GPA check is therefore essential, since grade errors were the most frequent error type for every model tested.

GPAs are matched to within 0.01, so that a figure printed in truncated form (2.80 for 2.807) still passes. A discrepancy is always a reason to re-examine the record, never a reason to alter a value.

The free spreadsheet checker

A spreadsheet without macros performs all three checks. The user pastes in the extracted course lines and enters the printed totals and, where one is printed, the grading key; each total and each term then displays either OK or CHECK. In testing, the spreadsheet detected all three AI errors planted in the fictional specimens:

- A grade incorrect by a modifier: the credit total passed, and the term GPA displayed CHECK.
- An omitted term: the overall total and that term displayed CHECK.
- A points value in the grade field: the points total and the average displayed CHECK.

What the file itself can tell you

Every PDF records the program that created it and the dates on which it was created and last modified; these details are read from the file itself, not by an AI tool. A file that has been re-saved in an editing program, or modified after its creation, warrants closer examination. Because such details can be erased, and because innocent re-saving also alters them, a flag of this kind never constitutes proof in itself.

A demonstration: one altered grade

On fictional Specimen 03, one printed grade (C+) was covered with a white box and A− typed over it. Although the alteration is invisible to the eye, three checks detect it:

1. **File details:** the PDF now names a different program and was saved again after its creation.
2. **Hidden text:** the original C+ remains in the file beneath the new A−.
3. **Arithmetic:** the printed semester GPA remains 3.00, whereas the visible grades yield 3.16.

The arithmetic check in this case is the same check that detects AI copying errors, and it applies equally to scanned documents. A discrepancy that the tool did not cause may reflect an error by the issuing institution or an altered document; it is therefore noted or queried with the issuer, but never corrected.

What to check for

An AI tool need not be perfect, because errors are caught by the check. The evaluator's role thus becomes that of the checker: AI drafts the record, but it does not replace the evaluator. The following list sets out the errors observed, in the order in which a checker should look for them:

1. **Every grade against the page.** The most frequent error was an omitted plus or minus, for example A− copied as A.
2. **Grades that appear too favorable.** Misread grades were higher than the printed grade in about four cases out of five.
3. **The recalculated GPA against the printed GPA.** A discrepancy identifies the lines to be re-read; this check detected every inflated record in the study.
4. **More than the credit total.** Every AI tool copied the printed credit total correctly, even when 40% of its lines were incorrect, so a matching total establishes little.
5. **Every term and page.** One prompt caused the tool to omit entire blocks of courses without notice.
6. **Figures that are not printed.** Any total or sum in the record must be visible on the page.
7. **Grades expressed in words** ("bestanden", "pass"). These were the fields most frequently left blank.
8. **Every row, on every occasion.** Two grades exchanged between rows pass every total and the GPA. A clean record took almost as long to check as a flawed one, so a brief inspection cannot distinguish a good record from a poor one.

Notes that an AI tool adds about a transcript are best treated as prompts for inspection rather than as statements of fact. In one instance, a note asserted that course codes were printed in reverse, whereas the page showed them printed normally.

The checker can make the same mistakes as the tool. In one timed check, the staff member replaced the transcript's final mean mark, which the tool had copied correctly, with a year mean printed higher on the page. Printed totals and averages should therefore be confirmed against the last figure on the transcript, by the spreadsheet as well as by the checker.

Using it with any tool

The method is a procedure rather than a piece of software. It combines a plain-text prompt, a fixed reply structure, arithmetic checks and a person who makes the decisions, and any AI tool capable of reading PDFs can perform the first step.

The TYPE method is shown in full on the next page.

Under this arrangement, AI performs the transcription, arithmetic performs the checking, and the evaluator makes every judgment.

Level	Suits	How
1. By hand	Small offices with no IT staff	Paste the prompt into a business AI chat tool, attach the transcript, ask for the course table, paste it into the free spreadsheet checker
2. Automated	Offices with Microsoft 365, Google Workspace or similar	Build the same steps as an automated flow: extract, check, flag, route to an evaluator
3. Custom	Offices with developers	Call an AI model directly and run the checks in code

This study tested Microsoft AI Builder and the ChatGPT, Claude and Gemini chat applications. With the v1.1 prompt, the chat applications copied between 98.2% and 99.6% of the fictional lines and Microsoft AI Builder (GPT-4.1 mini) copied 96.1%, yet the same tool running a larger model achieved only about 61%. Offices should therefore test the exact model and tool they intend to use. Other tools have not been tested, although the method itself does not depend on any particular tool.

Using a chat tool

1. Open a new chat for each transcript, with memory or personalization disabled.
2. Paste the prompt text and attach the transcript PDF in a single message, with no other content.
3. Request the course lines as a table, and copy it exactly from the reply.
4. Paste the table into the checker, enter the printed totals, and review the result.

The TYPE method: five steps

T Typist prompt Y Your own test first P Printed-totals check E Evaluator decides

Y Before you start: test your tool on the free made-up transcripts, and retest after any change.

▼

T AI 1. Extract

The tool copies every course line and every printed total, exactly as printed, into fixed fields. Typist, not interpreter: no converting, no guessing.

▼

P Arithmetic 2. Check

Add up the extracted lines and compare them with the printed figures:

- total credits
- points totals (credits × mark)
- GPAs, only where the transcript prints its own grade points

▼

P AI or evaluator 3. Re-read once

For each block that does not match: one narrow re-read of that block only. Accept it only if it now matches the printed figures. Done by hand, the evaluator re-reads the block against the transcript.

Step 3 happens at most once per block. No loops.

▼

P Arithmetic 4. Flag

Blocks that still do not match are flagged, with both numbers, so the evaluator sees them first.

▼

E Evaluator 5. Decide

The evaluator checks every line against the transcript, clears or confirms each flag, and does the evaluation. The tool's output is a first draft, never an evaluation.

The one rule that never bends: never change a value to make the numbers agree. Re-read, then flag. A mismatch the tool did not cause (a school's own error, or an altered document) is noted or queried with the issuer, never corrected.

Who does what: AI = copying only · Arithmetic = a spreadsheet, a flow or a script, never the AI tool · Evaluator = every judgment. Keep real applicant data in tools covered by a data-protection agreement.

5. Check every line against the transcript, re-reading first any block marked CHECK.

Chat applications generally do not permit the temperature to be set, so the same transcript may produce a slightly different reply if processed again.

Protect applicant data

Real applicant transcripts should never be submitted to personal or free AI accounts. The providers of such accounts may retain chats and use them to improve their models, and the office has no agreement with them covering student records. Business versions should be used instead, with contracts that cover data protection and exclude training on the office's data, and their compliance with FERPA, GDPR or the office's own obligations should be confirmed. Any new tool should first be tested on the fictional specimens.

From prompt to agent (future work)

A natural extension is to have software perform the checks and a single re-read automatically. That design is described here, although it has not yet been tested.

A prompt reads once and responds once, whereas an agent works in stages and can use tools along the way. For transcript extraction, an agent would:

1. Extract the transcript under the same typist rules.
2. Perform the credit, points and GPA checks in code, never through the tool itself.
3. Return each failing block to the tool once, with a narrow instruction to re-read that block only.
4. Accept the re-read only if it then matches the printed figures.
5. Flag any block that still fails, showing both figures, for the evaluator's first attention.
6. Pass the record to an evaluator, who checks every line and makes every judgment.

One rule admits no exception: the agent may re-read, but it must never alter a value merely to make the totals agree. An agent instructed to "make it add up" could change a grade to fit, which would constitute fabrication.

In summary, AI transcribes, code checks, and the evaluator decides. Grade conversion and equivalency remain with the evaluator and the office's own reference tables, never with AI.

Limits

This is a small, practical study; its findings indicate what to check rather than how every tool will perform. Its principal limitations are as follows:

- **Small sample.** The study used five real and ten fictional transcripts, with one run per combination of AI tool, prompt and transcript. Differences of one or two lines may be due to chance.
- **Construction of the keys.** The answer keys were prepared by the same team that wrote the prompts and conducted the tests. The real-transcript key was verified against every figure printed on the transcripts, but 32 lines could not be verified against any printed figure, and neither key was independently re-checked.
- **Shared basis for tuning and retesting.** Prompt versions 1.2 and 1.2b were derived from failures on the five real transcripts, and the fictional transcripts were designed from the same failures.
- **Partial record.** The Liberian transcript was a single page of a multi-page record.
- **Rapid change in AI tools.** The results apply to the named models, applications and settings on the dates given, and do not constitute a lasting ranking. The default models on the chat applications' free plans may change at any time.
- **Repeat runs not measured.** Temperature 0 in Microsoft AI Builder reduces variation between runs but does not eliminate it, and the chat applications offer no equivalent setting.
- **Timing by a single checker.** Typing was timed on one transcript and checking on a small number of records, all by one experienced staff member who was aware that she was being timed. Under the pressure of a normal workload, checking may be faster and less thorough.
- **Cleaner fictional transcripts.** Computer-generated scans and orderly layouts may favor AI tools relative to real records, and the regional transcripts are constructions based on common conventions rather than samples of any country's documents.

Confounding of layout and transcript. Because each layout was tested on a single transcript, the effect of layout cannot be separated from the document's country, scan quality, length or grading system. Whether layout itself determines accuracy is therefore a question for future research, which could isolate its effect by printing the same courses in different layouts and processing them with the same models.

Conflict of interest. The author is employed by Veri Eval, a foreign credential evaluation company that uses Microsoft AI Builder in its evaluation system and that may benefit from findings concerning AI-assisted evaluation. Veri Eval designed the test set,

conducted the tests and scored the results. To permit independent verification, the fictional transcripts, answer keys, prompts and scoring rules are published with this paper (*Free resources*).

Checklist for any office

Before an office relies on an AI tool to transcribe transcripts, the following steps are recommended.

Before you start

- [ ] Choose a business AI tool whose contract protects applicant data and excludes training on it.
- [ ] Use a typist prompt: copy only what is printed, never convert or guess, flag anything unclear, and never calculate in the fields.
- [ ] Test the tool on the free fictional specimens and score it against their keys.
- [ ] Set pass marks before testing: lines copied exactly, totals matched, no invented courses, and time saved.

On every transcript

- [ ] Compare the extracted credits with every printed credit total.
- [ ] Where marks and points are printed, compare the points totals.
- [ ] Where the transcript prints its own grade points, recompute and compare the GPAs.
- [ ] Re-read every block that does not match. Never change a value to make the totals agree.
- [ ] Check every line against the transcript in person; a passing check does not verify titles or codes.
- [ ] Note or query any discrepancy that the tool did not cause; never correct the transcript.

Over time

- [ ] Re-test whenever the model or the prompt changes.
- [ ] Keep a record of failed checks to identify where the tool performs poorly.
- [ ] Keep grade conversion and the evaluation itself with the office's evaluators.

Free resources

All materials needed for an office to test its own AI tool are freely available from verieval.com/research/test-kit under a CC BY 4.0 license.

Resource	What it is
Ten complete fictional transcripts (test set v2.0), scanned and clean, plus five short teaching samples	Scanned-look and clean PDFs, each reproducing one layout problem, marked "SPECIMEN — FICTIONAL"
Answer keys	For each specimen, in the prompt's JSON reply shape and as a readable table
Extraction prompt, version 1.2	The full instruction text, ready to copy (version 1.1 and the changes are in Appendix A)
Spreadsheet checker	Credit, points and GPA checks with no macros, plus a worked example with one planted AI error
Method diagram	The five steps on one page
Tamper demonstration	One altered grade and the three checks that reveal it
Scoring scripts	The scripts that build the specimens, verify them, and score AI replies against the keys

The transcripts are entirely fictional, with invented institutions, students and identification numbers, and must never be presented as genuine academic records.

Terms of use: CC BY 4.0. Share and adapt freely, with credit to Veri Eval LLC.

References

Sources consulted 31 October 2026. Details marked [to confirm] are completed before publication.

1. Association of International Credential Evaluators. (2024, November). *Use of artificial intelligence in credential evaluation: A principled approach*. Link

2. [Authors to confirm]. (2025). Automating academic transcript evaluation: A comparative study of OCR techniques for course and grade evaluation. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25)*. https://doi.org/10.1145/3698205.3733958
3. Bracken, D. K., & Colic, S. (2025). *Automating structured information extraction from images of academic transcripts using machine learning*. American Society for Engineering Education Annual Conference. Link
4. Council of Europe. (2026, May). *Situational analysis report on AI in qualification recognition* [title to confirm]. Link
5. DegreeSight. (2026, March 17). *How does DegreeSight handle transcript OCR for universities?* Link
6. Lowry, T. ([date to confirm]). *Using AI to evaluate undergraduate transcripts*. Law School Admission Council. Link
7. Microsoft. (2026). *What's new in Copilot Studio*. Microsoft Learn. Link
8. RTI International. ([date to confirm]). *How AI changed the way we code 1.5 million college courses*. Link
9. Schubert, E., & Bießmann, F. (2026). *Evaluating structured information extraction with open models in a high risk public sector application* [Preprint]. arXiv. https://arxiv.org/abs/2608.18289
10. TAICEP. (2025, October 14). *Position paper on the use of AI in credential evaluation*. Link
11. [Presenter to confirm]. (2025, October 23). *"Why hire humans at all?" A.I. tools for credential evaluation* [Conference session]. TAICEP 2025 Annual Conference, Cavtat, Croatia. Link
12. Tondro Consulting. ([date to confirm]). *Automated transfer credit evaluation in minutes: What we showed at AACRAO Annual*. Link
13. Torkian, B., & Zhou, J. (2026). *A multi-agent AI system for automated high school transcript processing: Collaborative document analysis at scale* [Preprint]. arXiv. https://arxiv.org/abs/2606.13916
14. Verionce. ([date to confirm]). *Using AI technology to enhance quality control of credential evaluations*. Link
15. Wattanacheep, B., Komkris, P., & Heymun, I. (2025). An OCR-based framework for automated verification of Thai and English academic transcripts. In *2025 International Computer Science and Engineering Conference (ICSEC)*. https://doi.org/10.1109/icsec67360.2025.11298018

Appendix A: The extraction prompt

A1. Version 1.1, as tested

The following is the exact instruction text used in every AI Builder run reported in this paper; the transcript PDF was attached as the document input.

```
You are a transcript extraction engine for a credential evaluation service.
Your only job is to transcribe what is printed on the document into
structured
JSON. You are a typist, not an interpreter.

RULES — follow all of them:

1. Extract ONLY what is visibly printed. Never infer, estimate, complete, or
   invent any value. A wrong guess is worse than an empty field.
2. If a value is illegible, ambiguous, cut off, or missing, set it to null,
   set "uncertain": true on that transcript course, and add an entry to
   "extraction_notes" saying what the problem is and where it is.
3. Copy grades EXACTLY as printed ("A", "B+", "70", "5.0", "AB", "P").
   Do NOT convert, translate, normalize, or map them to any other scale.
4. Copy credit/unit values EXACTLY as printed, as strings ("3", "4.0", "6").
   Do NOT compute, sum, or derive any number yourself. If the document prints
   its own totals or averages, copy them into "printed_totals" — never
   calculate your own.
5. Transcribe course names as printed, in the original language. Do not
   translate them.
6. Process every page. Keep courses in printed order. Record the page number
   and line position for every course. Copy the course code exactly as
printed
   into "course_code" (null where none is printed), and the term heading the
   course sits under ("2009/2010 First Semester", "Year 2") into
   "session_or_semester" (null where none is printed).
7. If a transcript course carries a repeat or carryover marker (for example
"R",
   "repeat", "carried over"), copy that marker into "repeat_indicator".
   Do not decide what it means.
8. If the document is substantially unreadable (poor scan, unrecognizable
   layout), set "unreadable": true, extract whatever is safely readable, and
   explain the problem in "extraction_notes".
9. If a grading-scale legend or key is printed anywhere on the transcript,
   copy its text verbatim into "grading_scale_hint".
10. Output ONLY the JSON object in the shape below. No prose, no commentary,
```

```
    no markdown fences, nothing before or after the JSON.

OUTPUT SHAPE (use exactly these keys):

{
  "applicant_name": "string or null",
  "institution_name": "string or null",
  "program_or_award": "string or null",
  "credential_dates": { "start": "string or null", "end_or_award": "string or
null" },
  "grading_scale_hint": "string or null",
  "printed_totals": { "total_credits": "string or null", "cgpa_or_average":
"string or null" },
  "language": "string or null",
  "courses": [
    {
      "page": 1,
      "line": 1,
      "session_or_semester": "string or null",
      "course_code": "string or null",
      "course_name": "string",
      "native_credits": "string or null",
      "native_grade": "string or null",
      "repeat_indicator": "string or null",
      "uncertain": false
    }
  ],
  "extraction_notes": [
    { "page": 1, "issue": "string" }
  ],
  "unreadable": false
}
```

A2. What version 1.2 changes

Version 1.2 retains all of the above, including the output structure, and makes the changes and additions listed below. The full text of version 1.2 is included in the free download kit.

Rule	Version 1.1	Version 1.2 adds or changes
3. Grades	Copy grades exactly; do not convert.	Adds: "native_grade" is the mark or letter printed for that course. It is NEVER a points, grade-points, quality-points, weighted or "GPA" column (a number that equals credits multiplied by the mark). If you cannot tell which column is the grade, set "uncertain": true and say so in "extraction_notes".
4. Credits and totals	"Do NOT compute, sum, or derive any number yourself."	Changes that sentence to: "Never put a number you calculated into any field." Adds: If a total or average is printed more than once (on every page, term or year), put the FINAL cumulative figure in "printed_totals" and say in "extraction_notes" which page it came from.
6. Every page, in order	Process every page in printed order, with page, line, code and term.	Adds: Many transcripts print courses in several blocks SIDE BY SIDE (terms, years, or the left and right halves of a page). Read one block completely, top to bottom, before starting the next; go left to right, then down the page. Take the credits and the grade from the SAME row as the course name. Never skip a block. Where a block prints its own credit total, check that your lines for that block add up to it; if they do not, re-read the block, and if they still do not, say so in "extraction_notes".

Rule	Version 1.1	Version 1.2 adds or changes
11. Final check	Not in version 1.1.	New: FINAL CHECK, before you output: add up "native_credits" for every course that has a number there. If "printed_totals.total_credits" is printed and your sum is different, add ONE entry to "extraction_notes": { "page": <last page>, "issue": "Credit check: courses add to <your sum>; printed total is <printed total>." } This is the ONLY place a calculated number may appear. Never change, add or remove a course to make the numbers agree.

A3. Version 1.2b

Version 1.2b consists of version 1.1 with two of the four version 1.2 changes added verbatim: the rules on repeated totals and on blocks arranged side by side. It omits the grade-column rule and the final credit check, and it was used only for this study, never in production.

Rule	Version 1.1	Version 1.2b
4. Credits and totals	"Do NOT compute, sum, or derive any number yourself. If the document prints its own totals or averages, copy them into "printed_totals" — never calculate your own."	"Never put a number you calculated into any field. If the document prints its own totals or averages, copy them into "printed_totals" — never calculate your own. If a total or average is printed more than once (on every page, term or year), put the FINAL cumulative figure in "printed_totals" and say in "extraction_notes" which page it came from."

Rule	Version 1.1	Version 1.2b
6. Pages and order	Ends after the course code and term heading sentences.	Adds: "Many transcripts print courses in several blocks SIDE BY SIDE (terms, years, or the left and right halves of a page). Read one block completely, top to bottom, before starting the next; go left to right, then down the page. Take the credits and the grade from the SAME row as the course name. Never skip a block. Where a block prints its own credit total, check that your lines for that block add up to it; if they do not, re-read the block, and if they still do not, say so in "extraction_notes"."

All other content, including the reply structure, is identical to version 1.1.

Appendix B: The five teaching samples

Specimen	Layout	Lines	What it tests
01 Single column (control)	One page, marks out of 100, printed average 65.00, grading key	20	Baseline: should be copied perfectly; average copied exactly
02 Two columns, "Page 2 of 4"	Letter grades with quality points; a term split across columns; page total 58 beside cumulative total 141	22	Reading order; whole-record total taken for the page; credit-check note
03 Term blocks side by side	Four terms in a 2 × 2 grid; 6 rows per pair differ only by a modifier; cumulative GPA printed cut short	40	Modifier errors; credit check passes while GPA check fails

Specimen	Layout	Lines	What it tests
04 Four year-columns	Landscape; term totals and GPAs; fractional credits; a 3-course Supplementary Term in small type	43	Skipped term; repeated values; stopping early; fractions
05 Mark and points columns	Two pages; mark beside points (credits × mark); carried and brought forward totals	40	Points copied as grades; repeated totals

A separate script recomputed every printed total, average and GPA on the specimens from the answer keys and confirmed that every key line appears on the page.

Appendix C: Detailed results

This appendix holds the detailed tables summarized in the Results section.

C1. Real transcripts: what went wrong (prompt v1.1)

Lines are counted once each; one line can have both a grade and a credit error.

	GPT-4.1 mini	GPT-4.1	GPT-5 chat
Lines wrong or missed	73	180	189
… with a grade error	58	152	124
… with a credit error	9	58	49
… missed entirely	8	10	55
Lines the model flagged uncertain	1	0	1
Runs that noted their own credit mismatch	0 of 5	0 of 5	0 of 5

C2. Real transcripts: v1.1 and v1.2 by transcript

Transcript	v1.1 lines right	v1.2 lines right
Liberia (two columns)	21 / 21	21 / 21

Transcript	v1.1 lines right	v1.2 lines right
Kenya (one column)	65 / 65	65 / 65
Pakistan (side by side)	35 / 52	36 / 52
Bangladesh (four year-columns)	35 / 73	37 / 73
Afghanistan (mark and points)	66 / 84	56 / 84
All five	**222 / 295**	**215 / 295**

C3. Header details

Header details were scored separately from course lines across all 125 replies: applicant name, institution, program, start and end dates, printed totals, language and grading key, giving 1,125 fields in total. Under the strict rule, 85.0% of fields matched the answer key; under the lenient rule, which accepts both language parts of a bilingual entry in any order and recognized language names written as codes ("fr" for French), 94.2% matched. Course lines, for comparison, were 89.0% right. Identity fields (name, institution, program and dates) were 90.2% right under the strict rule, and printed totals 91.2%. Most remaining errors fell into three patterns. First, on the bilingual Chinese (R3) and Arabic (R4) transcripts, most tools kept only the English half of the name, institution and program and dropped the other script. Second, GPT-4.1 mini frequently left the language field blank. Third, C2 prints no credit total, yet GPT-4.1 mini on v1.2 and v1.2b entered 64, a figure it had calculated by counting units; this breaks the copy-only rule and would mislead an evaluator who took it as printed.

C4. The 22 predictions

No.	Prediction	Result	Evidence
P1	GPT-4.1 mini on v1.1 copies fewer than 95% of lines right	Not held	539 of 561 (96.1%)
P2	C1 and C2 are the most accurate mirrors	Not held	C5 (97.2%) and C3 (92.0%) were the most accurate; C2 (75.8%) the least
P3	C4 is the least accurate mirror; a run skips its supplementary term	Not held	No run missed a C4 line; C4 was least accurate only within AI Builder runs

No.	Prediction	Result	Evidence
P4	GPT-4.1 and GPT-5 chat are less accurate than GPT-4.1 mini	Held	210 and 212 against 327 of 345
P5	No run invents a course	Held	No invented lines in 125 replies
P6	v1.2b gets more lines right than v1.1 on GPT-4.1 mini	Not held	531 against 539
P7	On C5, v1.2 copies points into the grade field	Not held	No such lines; every C5 run was fully right
P8	v1.2 scores below v1.1 on C5	Not held	Both 84 of 84
P9	v1.1 copies an earlier cumulative figure at least once	Not held	Every run copied the final figure on C1, C2 and C5
P10	Lost or added + and − are the most common grade error on C3 and C4	Held in part	Held on C4 (63 of 114); not on C3 (19 of 42)
P11	The credit check catches every reply that misses lines, and no reply whose only errors are grades	Held	6 of 6 caught; none of 20 grade-only replies flagged
P12	The GPA and term checks catch most grade errors the credit check misses	Held	229 of 333 errors (69%)
P13	The only false alarms come on R2 and C2	Not held	False alarms on C2 and R1; none on R2
P14	A chat application translates a course title or grade word	Held	Claude on v0 rendered the Arabic titles of R4 in English; ChatGPT on v0 rendered the German grading key of R5 in English
P15	A run changes comma decimals to points	Held	ChatGPT on v0 wrote the R5 grading key with points (1.0–1.5); course lines kept their commas
P16	The right-to-left layout of R4 puts values in the wrong fields	Not held	No field swaps; one credit misread on one line

No.	Prediction	Result	Evidence
P17	On R5, a run rounds the final grade or puts a number against a pass-only module	Not held	All kept 2,1; seven runs instead left the two pass-only grades blank
P18	GPT-4.1 mini copies the Chinese titles on R3 less accurately than the chat applications	Not held	GPT-4.1 mini made no title errors
P19	ChatGPT, Claude and Gemini on v1.2b each beat GPT-4.1 mini on v1.2b	Not held	Claude 345 and Gemini 342 did; ChatGPT (279) did not, against 316
P20	A chat-application reply adds text or formatting around the JSON	Held	ChatGPT replies carried formatting marks, possibly from the copy step
P21	Header fields are at least 90% right	Held in part	85.0% strict; 94.2% lenient
P22	AI plus checking is faster than typing on C2, C3 and C5	Held in part	Faster on C2 and C5; C3 was not checked; the smallest gap was on C2