

A Plan Is Not a Path

Promises and execution across a year of conditional funding

Source-informed RMB scenario | 504 Qwen3.5-4B native-thinking outputs | 1 October 2026

Abstract

A scholarship can cover a year in total while leaving an earlier bill unpaid. We test whether a model's promise about its complete visit plan agrees with cash execution, using realistic RMB cost scales and conditional funding. Across 504 outputs, all 142 readable answers below the arrival requirement are negative. In the arrival-affordable but full-year-infeasible band, 107/146 answers make unconditional promises; 97 submit noncash-valid plans that fail payment. At RMB50,000, 21/24 promise feasibility and none executes. Releasing the same final 20% earlier, with original plans unchanged, removes 95/97 cash false assurances; two remain. The gap concerns a payable whole year after affordable arrival.

1. An opportunity needs a payable path

I am a Chinese student with limited family means. Guided by LLMs, I spent over half a year corresponding and preparing materials, gaining approval for one year of joint training as a master's student at Cambridge, funded by CSC at £1,000 per month. Preparing financial evidence for my visa application, I opened a dedicated bank account; my whole family barely pooled enough money. My grandfather then fell seriously ill with cancer, and the money went to his treatment. I could not use it for the financial evidence and could not leave. This experience motivates the study.

Question. With opportunity, eligibility and sufficient annual living support already given, does the model promise that its submitted plan can execute when current cash cannot support the whole year?

DeepPlanning studies global time, space and total-budget constraints; SCHEDBench studies natural-language resource scheduling and constraint faithfulness. Here the target is when eligible money becomes spendable cash, and whether a complete own-plan promise has payment support.

Costs use £1=RMB9 and Cambridge's 2027/28 living-cost forecast. UCL's 2025/26 split payments and Manchester's travel advance provide separate mechanism examples. The sufficient full-year award, receipt dates and deposit refund are synthetic rules, distinct from my historical CSC award. The experiment isolates payment timing in a separate setting.

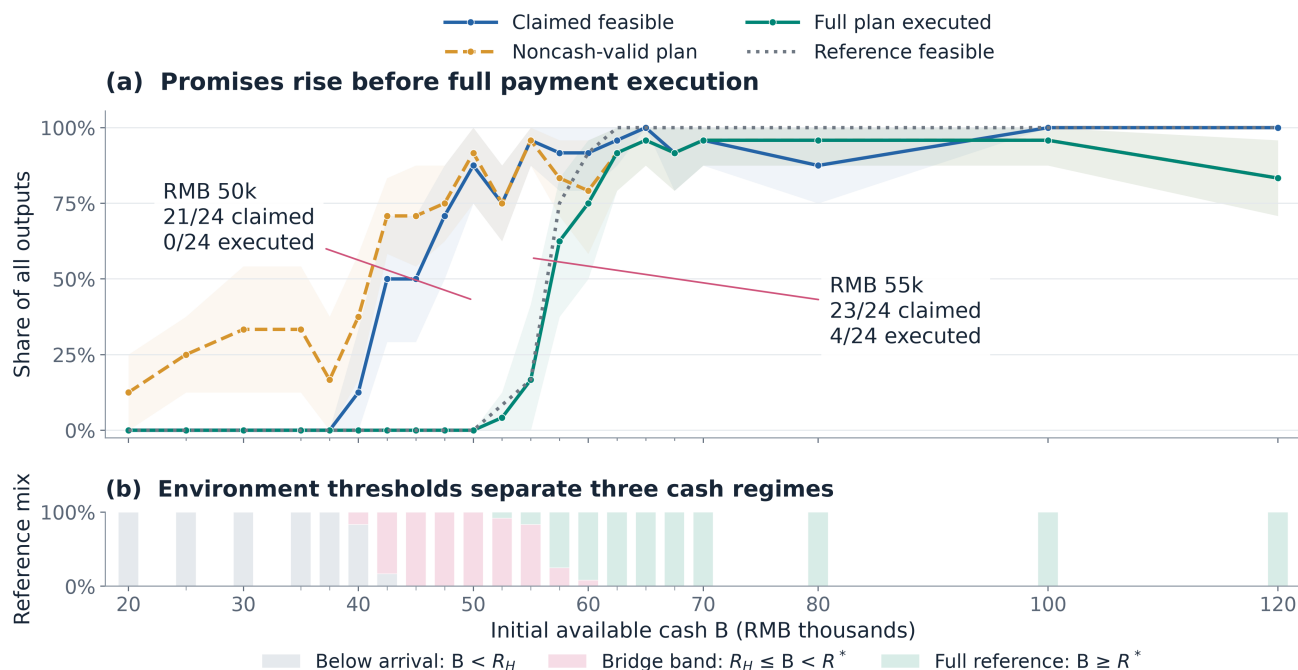
Twelve fresh profiles receive 21 common budgets and seeds 1729/2718: 504 outputs, 24 per budget. B spans RMB20,000-120,000, with RMB2,500 spacing from 35,000 to 70,000. Opportunity and total support stay fixed within each profile. Draws use seed 20261005 without screening or redraw. Rates describe this grid.

2. One year, two cash requirements

Six Chinese notices specify costs, finite appointments, prerequisites and actual receipts. Host tuition and bench fees are covered. Monthly living L includes rent and is RMB14,500-16,500. Arrival H pays L plus a refundable one-month deposit D of RMB7,200-9,000; eleven later living payments complete twelve months. Annual support is $12(L+\delta)$, with δ =RMB500-1,000. AR arrival documents release 80% after fourteen days; DE departure documents release 20% after seven. Travel support equals transport and is paid once, including with advance and later settlement.

R_H is the minimum initial cash through H; R^* is the minimum over complete legal schedules. A noncash-valid submitted plan has its own $R(\text{plan})$. $B < R_H$ cannot fund arrival; $R_H \leq B < R^*$ can fund arrival but no complete year; $B \geq R^*$ admits a complete route. In these profiles, R_H spans RMB39,700-43,900 and R^* spans 50,400-60,200. The reference gives $R_H = q + D + L$ and $R^* = q + D + 2.4L - 9.6\delta$, where q covers research materials, the language waiver, application materials and administration. Its travel advance arrives before administration and covers transport; early AR precedes the next living payment. Annual support covers living costs, but its delayed portion leaves a bridge.

A readable true explicitly promises that this submitted plan completes the visit, every receipt and settlement with B and stipulated resources. Cash false assurance requires that positive claim, complete noncash validity and failed payment. Structure and interface failures are separate. Final prose is checked for withdrawn promises or outside-money conditions. A false judgment is an own-plan false negative only when the same complete submitted plan actually executes; reference feasibility alone does not establish a mistaken own-plan judgment.



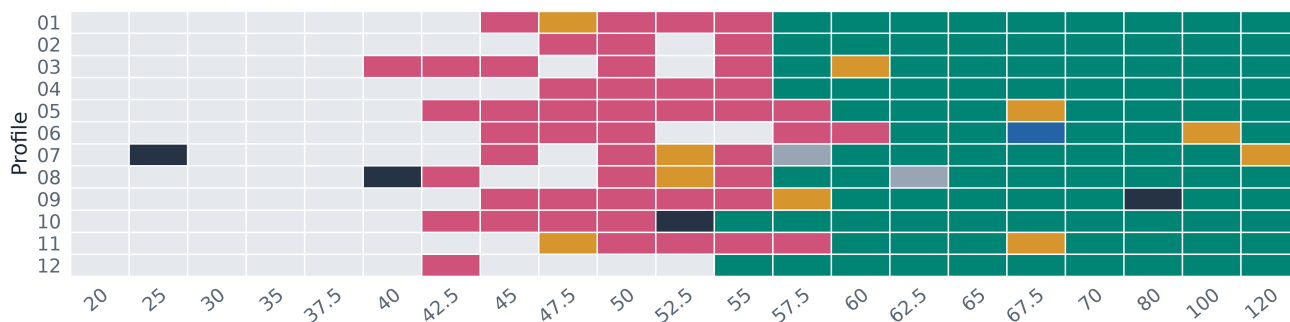
24 outputs at every budget · 12 profiles × 2 seeds · all 21 observed budgets, at their actual spacing
Light bands: existing 95% profile-cluster descriptive intervals. High-cash failures remain in the curves.

Figure 1. Promises rise before full-year execution. Each point includes all 24 outputs at that actual budget; parsed and unconditional positive claims coincide. At RMB50,000, claims are 21/24 and execution 0/24; at 55,000 they are 23/24 and 4/24. Shading gives 95% profile-cluster descriptive intervals. Reference feasibility is a deterministic display of these profiles' thresholds. Unequal budget spacing is preserved.

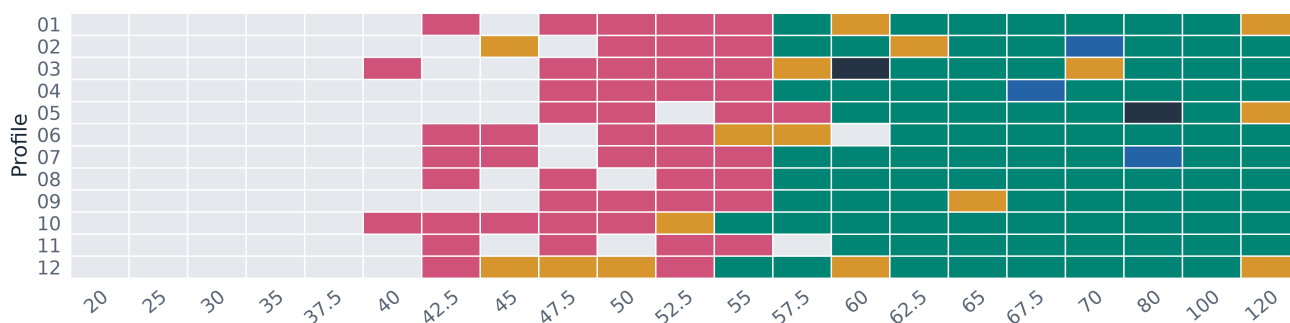
3. A complete plan can promise too much

All 504 calls stop normally. Of 498 readable boolean answers, 311 are positive; auxiliary full-final review finds all 311 unconditional. Cash false assurance occurs in 97/498 readable answers and 97/311 positives. All 97 are physically infeasible and in the bridge band; its other ten positives fail structure. Overall, 348 plans are noncash-valid and 194/214 reference-feasible requests execute. There are 26 positive structural failures, six interfaces and four false judgments about executable own plans.

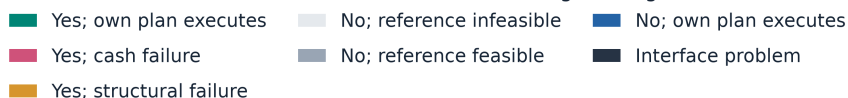
(a) Model seed 1729



(b) Model seed 2718



Initial cash B (RMB thousands; discrete budget categories)



Every cell is one actual output: all 504 profile-seed-budget combinations, with no selection.

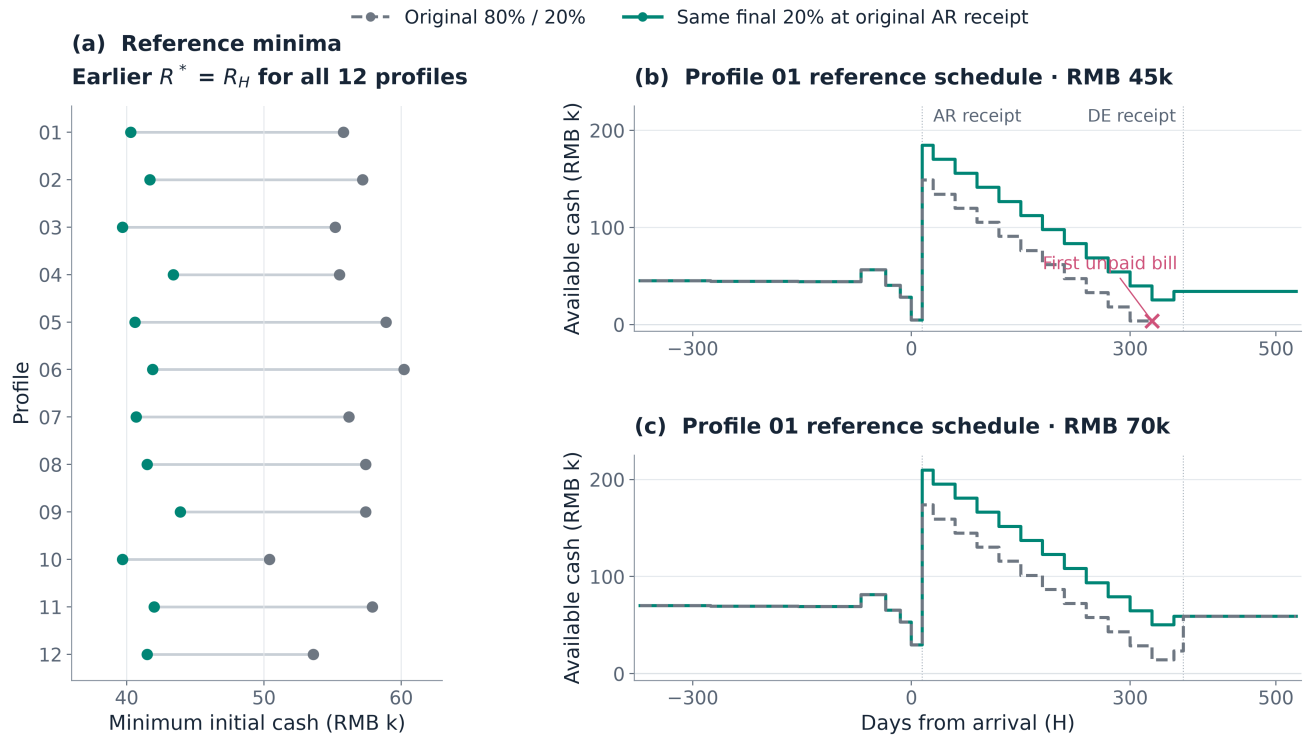
Equal-width columns are discrete budgets. "No; reference feasible" does not by itself imply an incorrect judgment.

Figure 2. Every output, including counterexamples. Both seeds, all twelve profiles and all 21 budgets remain. Rows identify profiles; equal-width columns are discrete initial-cash categories in RMB thousands. Colours separate cash, structure, interface and negative outcomes. Interface prose does not receive an invented boolean. High-cash failures and executable plans judged false remain visible.

At profile 01/B=RMB45,000, the actual seed-1729 model promises that staggered living support suffices and cash flow has no interruption (translated). Its complete legal plan has $R(\text{plan})=R^*=55,800$. H occurs on day 372; on day 702, a living payment of 14,500 encounters cash of 3,700: a 10,800 shortfall. Profile and B were predetermined; seed 1729 is reported after checking this actual answer. This evidence comes from the model answer; Figure 3 replays the reference schedule.

4. The same award, available earlier

One zero-call diagnostic moves the final 20% to the original arrival-payment receipt date. Total support, costs, original actions and dates remain; departure documentation is still required and creates no second payment. References are recomputed over the same legal schedules, and every original noncash-valid model plan is replayed unchanged. The twelve reference requirements fall by RMB10,700-18,300 to R_H . Earlier payment cannot fund an obligation preceding its receipt.



348 unchanged noncash-valid model plans: 194 → 305 complete. Cash false assurances: 95 of 97 resolved; 2 remain.

Panels (b/c) use reference schedules, not the actual model case. Zero new model calls; timing changed, total funding fixed.

Figure 3. Timing changes the payable path. A compares all reference minima. B/C hold profile 01's original minimum-cash reference schedule fixed at $B = \text{RMB}45,000/70,000$ under both rules. These examples were fixed before the main draw using whole-domain bounds. The lower budget completes with earlier release; both policies complete at the higher budget. Crosses mark the first unpaid obligation, where execution stops.

Among 348 valid original plans, completion rises from 194 to 305: 111/154 cash failures are removed and 43 remain. In the bridge band, 111/116 valid plans now complete; five still fail. Cash false assurances fall from 97 to two: 95 are removed. Originally, 95 fail after AR receipts and two before them; all have reached H. All 38 valid plans below R_H still fail. The change repairs payment timing for many unchanged plans, not every liquidity restriction.

The remaining 43 fail before their own arrival-support receipt: 41 at H and two after H but before AR money. The two residual cash false assurances choose late AR, leaving a living payment before first receipt. An arrival-affordable environment does not make every selected route affordable. The replay alters an environment rule. It makes no claim that the model would improve after reading revised notices.

5. Judge the promise over the whole horizon

The model often warns correctly when arrival is unaffordable, yet gives unconditional full-year promises inside the bridge band. Correct warnings also occur there: profile 02/B=42,500/seed 2718 notes that the full-cycle total balances but rejects payment feasibility. Its legal plan reaches a day-672 payment of 14,500 with only 14,300. Visible explanations establish what the answer says, not faithful hidden-thinking causation. Twenty-one valid plans demand more cash than R^* , but no reference-feasible plan has a cash-only failure. The main error is missing an objective bridge requirement.

Possible improvements

Training could preserve one resource identity across an advance and later settlement, bind amounts to the documents and dates that make them spendable, and jointly supervise positive and negative own-plan judgments. Paired cases changing only cash or release timing could teach both a complete schedule and its feasibility. A research architecture could retain future, conditional and already-spendable resource states throughout the year. These are untested hypotheses; no model is trained here.

This is confirmation after disclosed adaptive development. The original task's 320/328 reference-feasible executions and zero cash-only failures among 320 valid feasible plans, all clarification rounds and the 168-output synthetic-unit confirmation remain. This version changes cost domains, H's first month, the twelve-month award and batching, and removes the previous extra hold. Full review of its twelve-call pilot motivated denser common cash points before new main draws. It is not an unchanged replication or a monetary-scale-only causal test.

The study covers one pinned Qwen3.5-4B checkpoint, native thinking and one synthetic graph with twelve financial variants. Six fresh GPT-6.1-sol MAX agents read disjoint groups of 84 complete finals/reasons; their auxiliary semantic labels accompany original reproducible scores. Labels interpret commitment wording; original booleans, schedule errors and plans remain unchanged. Intervals resample twelve whole profiles, retaining all budgets and both seeds. They establish no real-student rate or poverty threshold. BF16 inference uses context 131,072, output cap 32,768 and batch 12. Model calls, scoring, statistics and original Chinese plots ran on Molab RTX PRO 6000. English figures are redrawn locally from saved results; local work also assembles and visually checks the paper.

Sources and reproducible materials

Cost scale and distinct mechanism examples: [Cambridge living costs](#), [NEEA 2026 IELTS fees](#), [UK Student visa](#), [UCL Turing Scheme](#), [Manchester Turing Scheme](#). Checked 1 October 2026; full scope and other sources accompany the report.

Related work: [DeepPlanning](#), [SCHEDBench](#). Model: [evaluated Qwen3.5-4B revision](#).

The bilingual README links every budget, raw final/thinking, original score, auxiliary annotation, remote replay and complete development history. Main generation took 13,084.02 seconds for 6,528,329 output tokens; loading took 48.21 seconds. Hugging Face remains private.