

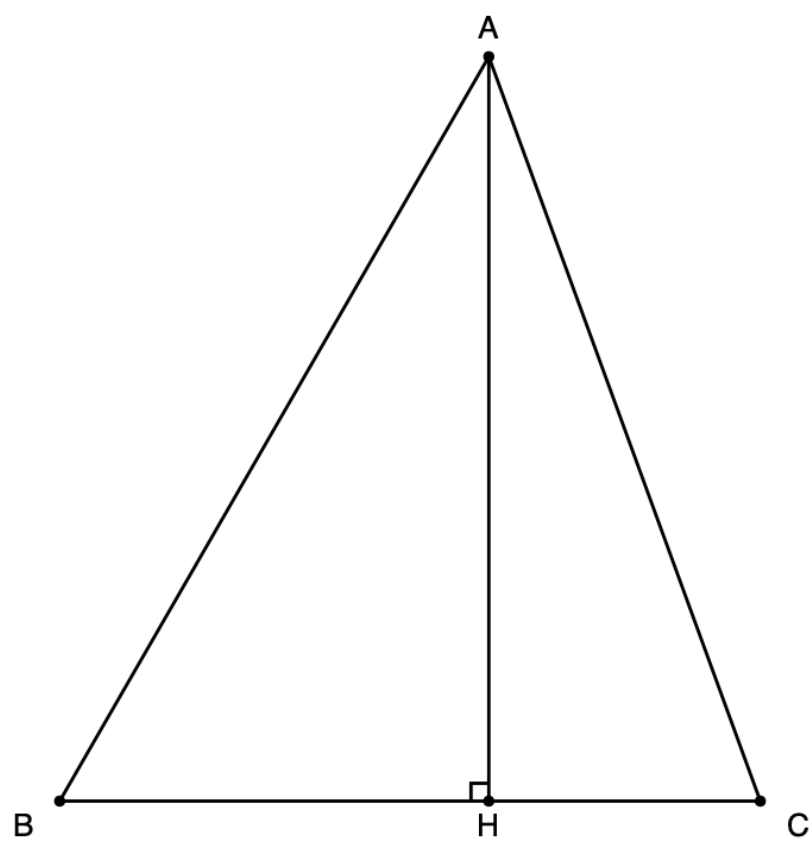
A geometry drawing can be graded exactly from its coordinates, with no judge. Graded this way, the scaffold matters more than the model.

1 The problem

Every geometry benchmark asks a model to read a diagram. None asks it to draw one, because grading a drawing has been the hard part.

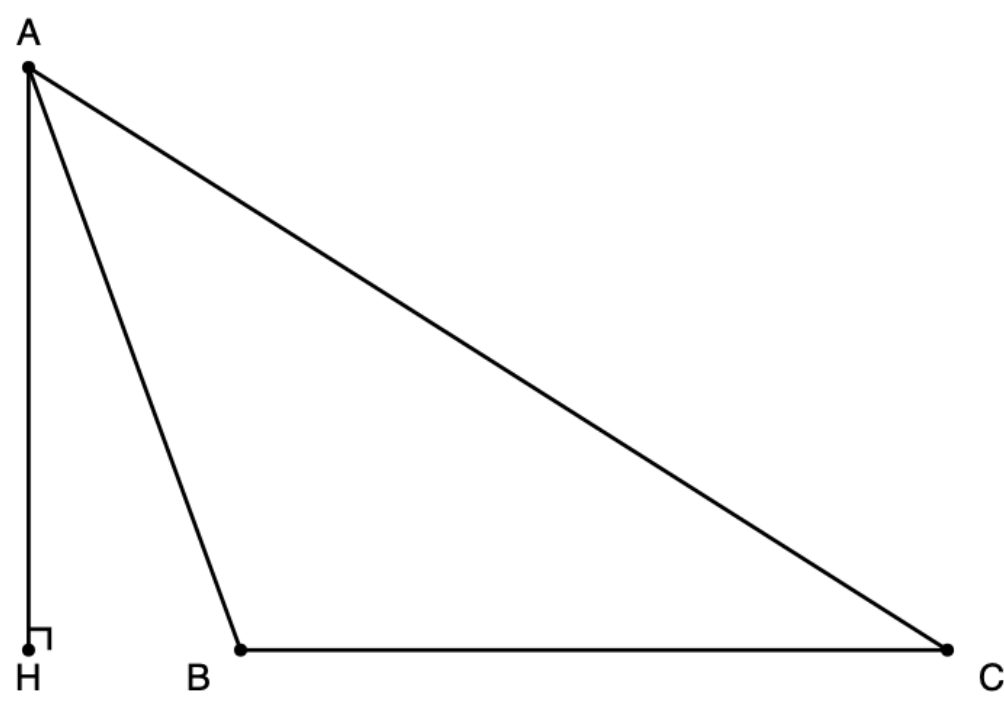
Both drawings below answer: “Draw an acute triangle ABC with 60° at B and 70° at C. Drop the altitude from A to BC, meeting it at H.”

PASS



60° at B, 70° at C, altitude from A, foot H on BC. Every check holds.

FAIL



Obtuse at B, so H lands outside BC. The right angle is still there. A person glancing at it would likely pass it.

Looking right is not the same as being right.

2 What is new

801

prompts, each paired with the checks a correct drawing must pass

15

checks, each a few lines of arithmetic. No human, no vision judge

6

frontier models. A typed scaffold lifts every one, by 2 to 25 points

0

correct drawings failed by the grader in a 200-output human audit

3 Code and data

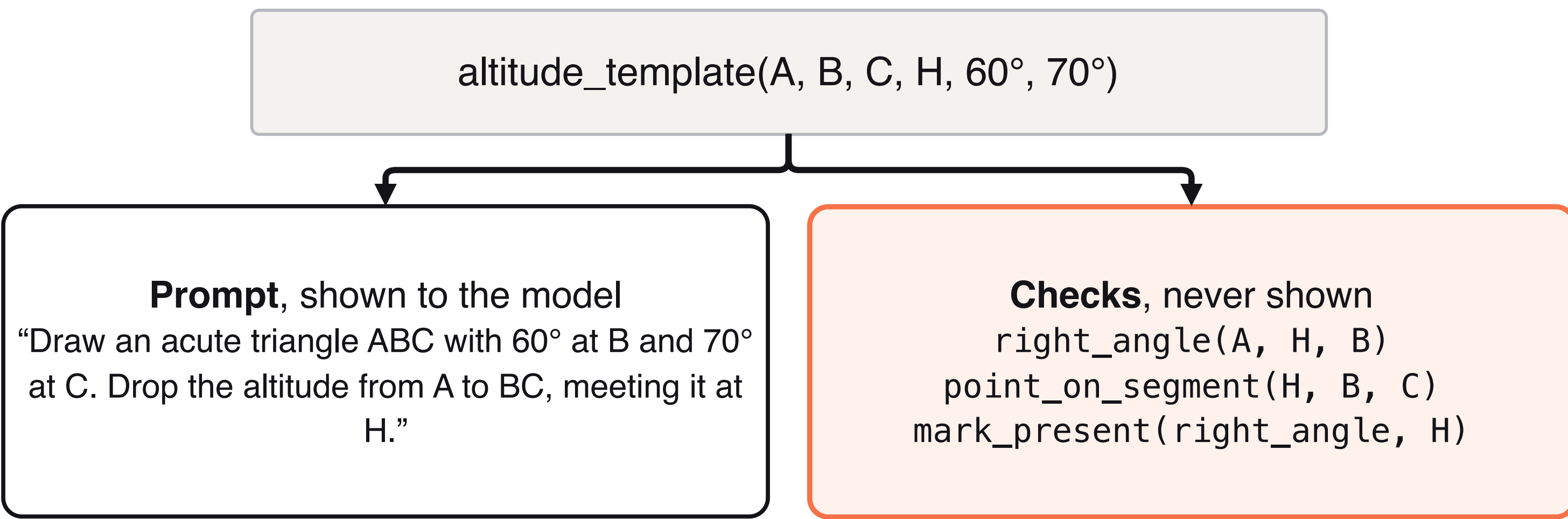
QR code
to add

Prompts, templates, grader, and leaderboard.

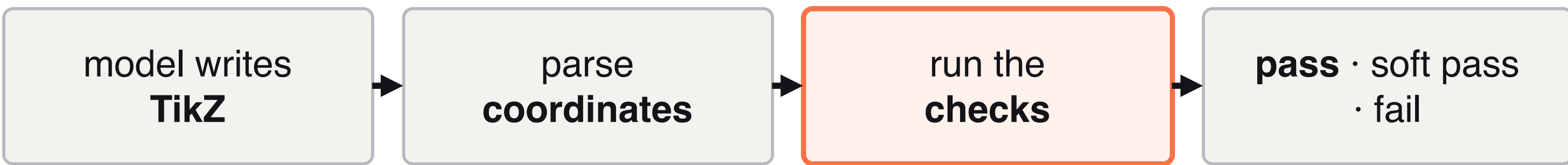
Release link to be added.

4 How a drawing is graded

One template writes both. A Python function emits the sentence the model sees and the checks the grader runs, from the same variables, so the two cannot disagree.

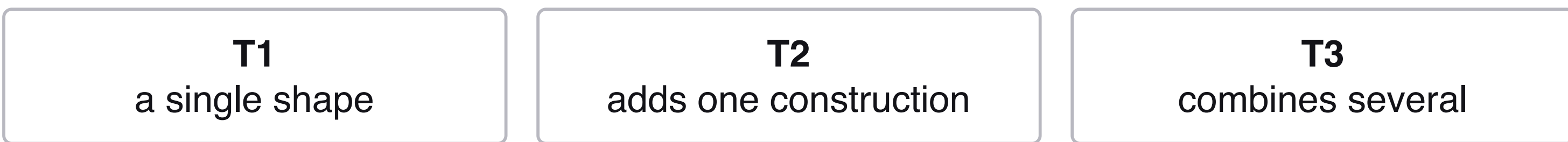


From drawing to verdict. The model writes TikZ. We parse the coordinates back out of it and run the checks on them.

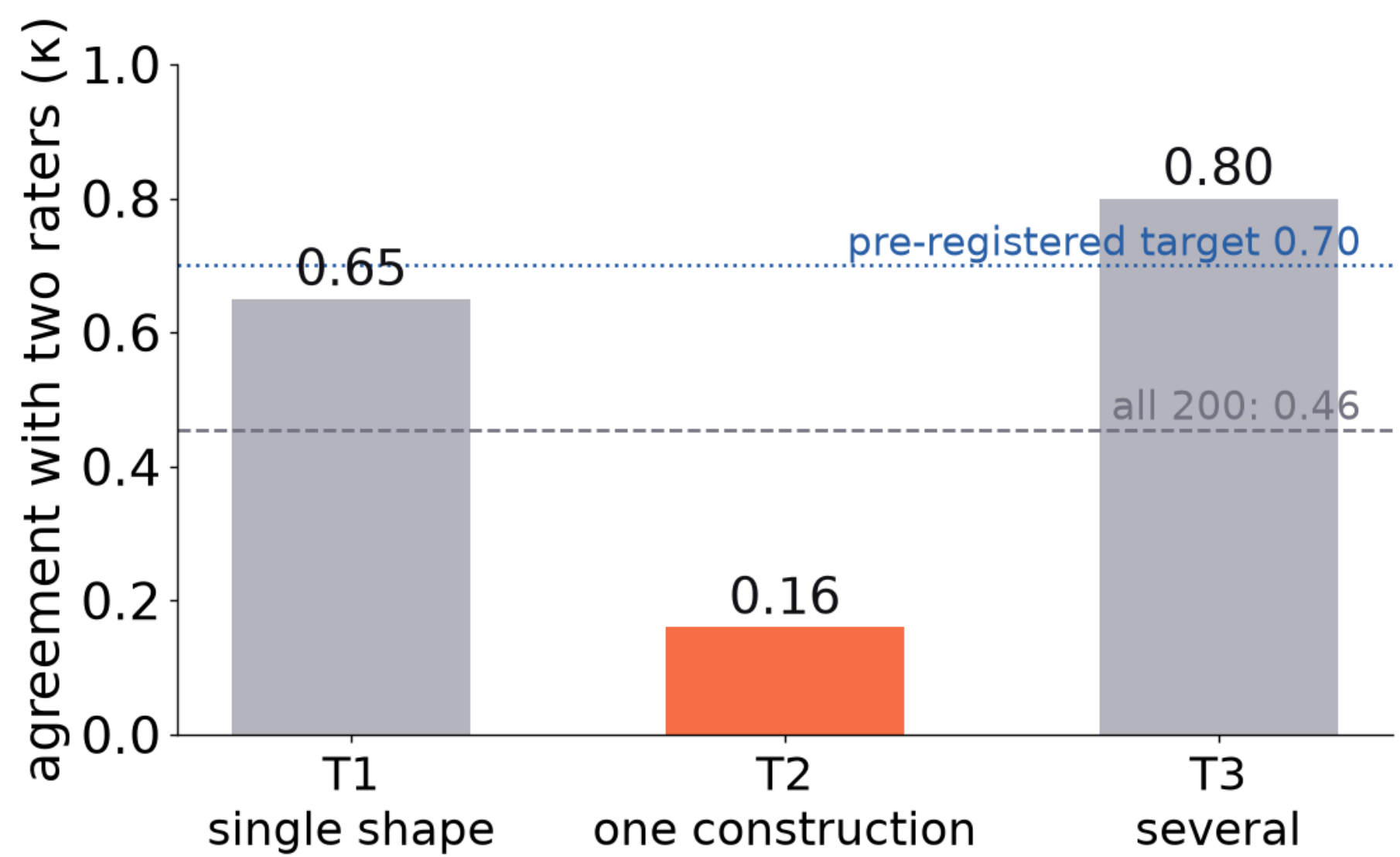


Fifteen checks such as right angle, midpoint, parallel, and point on segment, each arithmetic with a small tolerance. **A right angle is a dot product, not an opinion.**

600 prompts from 30 templates, 200 per tier, plus 201 from a K–12 curriculum.



5 Can the grader be trusted

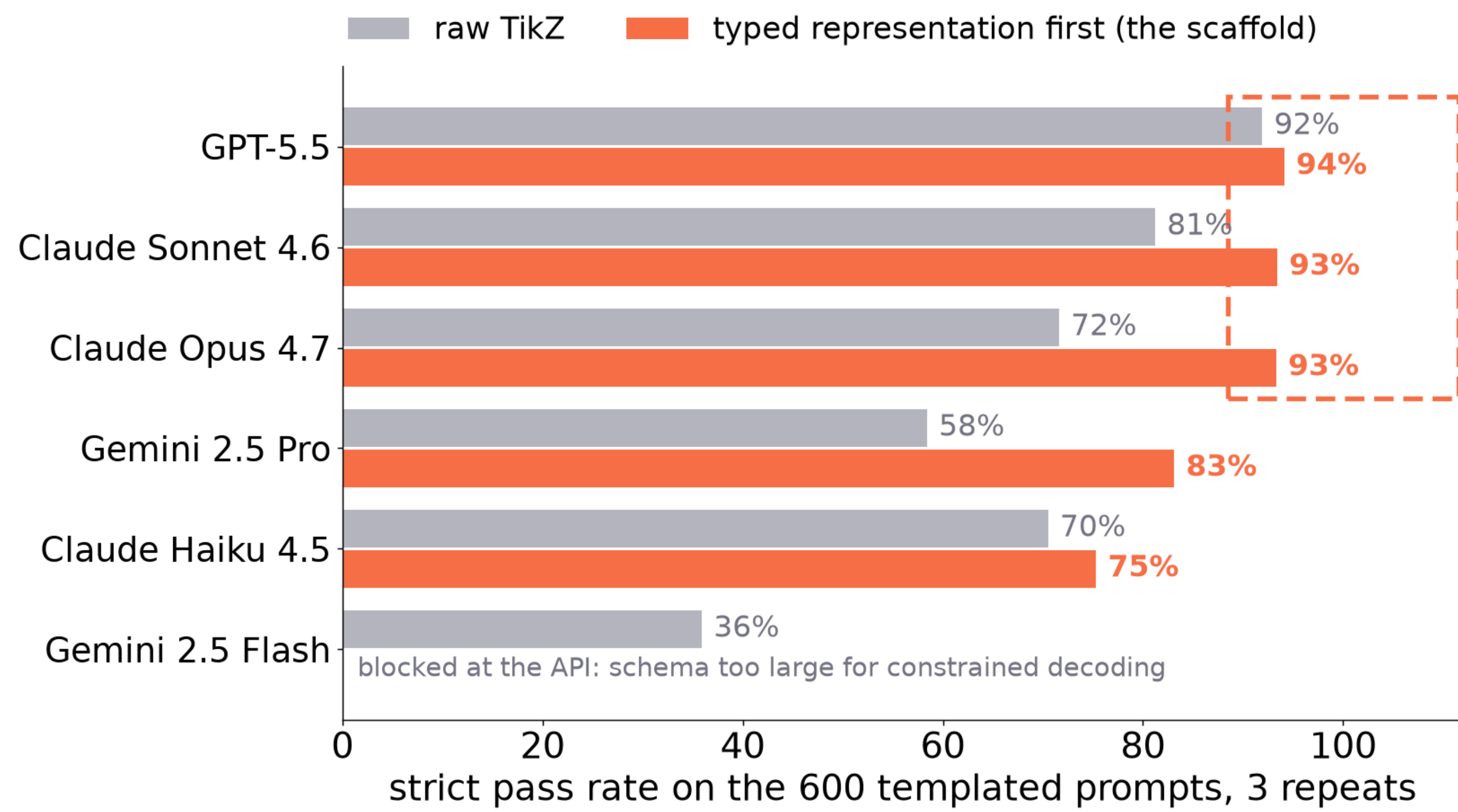


The disagreement sits in one tier, T2, where the templates check the construction but not the angles the prompt gave.

18%
of the grader's passes
overturned by two blind raters

0%
of its fails

6 What we found



Strict pass rate on 600 templated prompts, 3 repeats. Dashed box: three models within one point of each other at a 7× cost spread.

+2 to +25

The scaffold lifts every model. The weaker a model is at raw TikZ, the more it gains.

1 point

The frontier converges. GPT-5.5, Sonnet, and Opus finish within one point. Opus costs seven times more.

92 → 36%

Without the scaffold the same six models spread across 56 points.

Wrong drawings, not broken code. Most failures are clean, rendered diagrams with the wrong geometry. The grader catches them. A person glancing at the picture often would not.

Limits. The top models are near the ceiling on T1 and T2. On the curriculum prompts every score we ran roughly halved. The scaffold bundles format, compile path, and retries; we have not separated them.

Each requirement on a drawing can be checked exactly, from its coordinates.

Graded this way, the scaffold matters more than the model.