TensorTest · Research · November 2026

How close can frontier models get to an artist's 3D character?

We asked three frontier models to build the same three characters in Blender, then compared each build with the artist's own model. The best builds get the overall shape close. The faces, the joins between parts and the surface detail still differ, and those are what make each character look like itself.

TensorTest in collaboration with Atrium



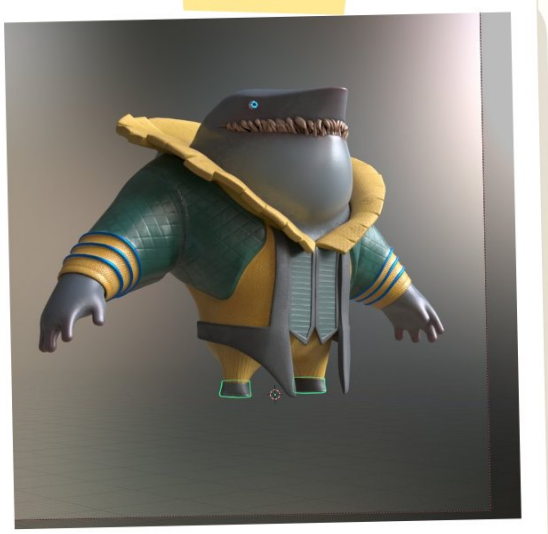
Lumi

A purple humanoid with molten orange hair inside a glass lamp, in short overalls.



Bess

An older horse with a long neck, a curly bun, red glasses and a utility vest.



Brick

A broad, squat shark in a green and gold jacket with a deep collar.

What we found

1. Two of the three models, GPT-6 Astra and Gemini 4 Argon, get the overall shape of Lumi and Bess close. Muse Wonder changes the proportions of all three characters.

2. Brick, a broad shark in a deep collar, is hard for every model. Even the closest builds miss the shape of its head.

3. In the closest builds, what remains is in the face, the joins between parts and the surface detail. Most of it is within 1% of the character's height, and it still changes how the character reads.

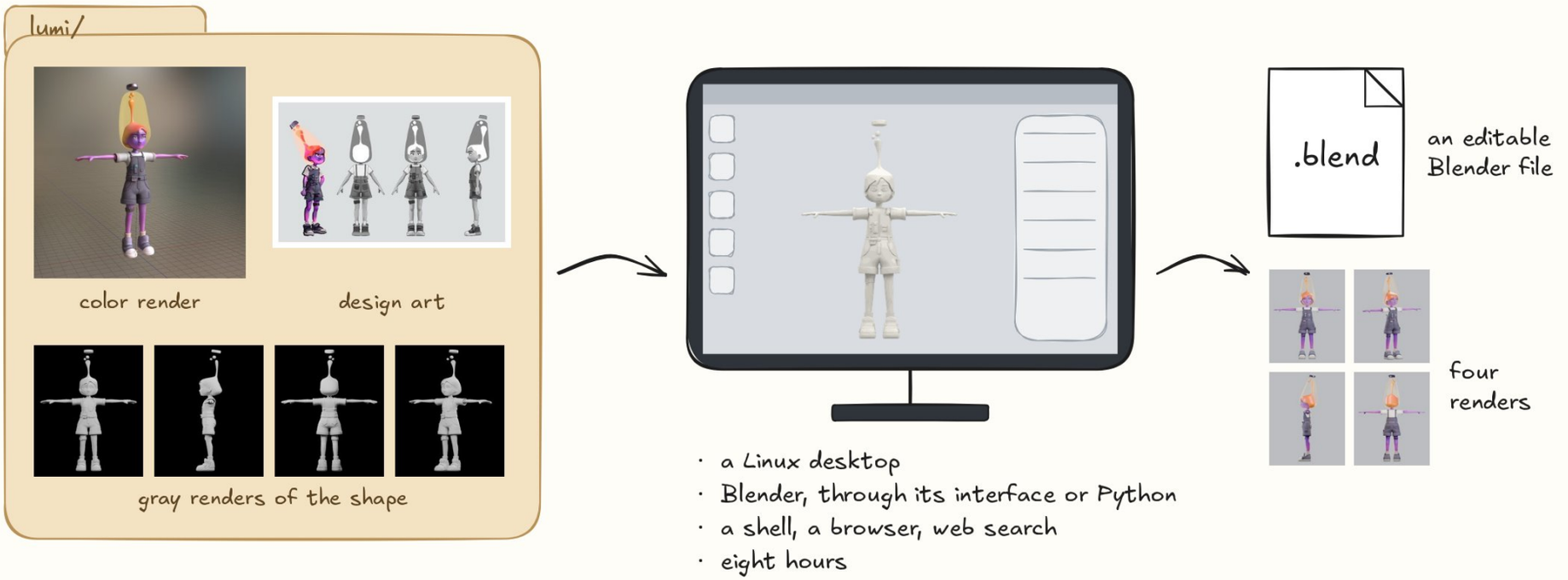
4. Whole-body scores barely register those differences. Grading a character model takes close views and an artist's eye.

01 The task

Nine builds from the same folder of images

Each model got one character and a folder of images: color and gray renders of the artist's finished model from the front, side, back and three-quarter view, closer views of the head, and some of the original design art. It worked in a Linux desktop with Blender, which it could drive through the interface or with Python, plus a shell, a browser and web search. It had eight hours to hand back an editable Blender file and four renders.

We ran GPT-6 Astra, Muse Wonder and Gemini 4 Argon, each at its highest reasoning setting, once on each character.





The artist's 3D model was never in the folder. We used it only to compare.

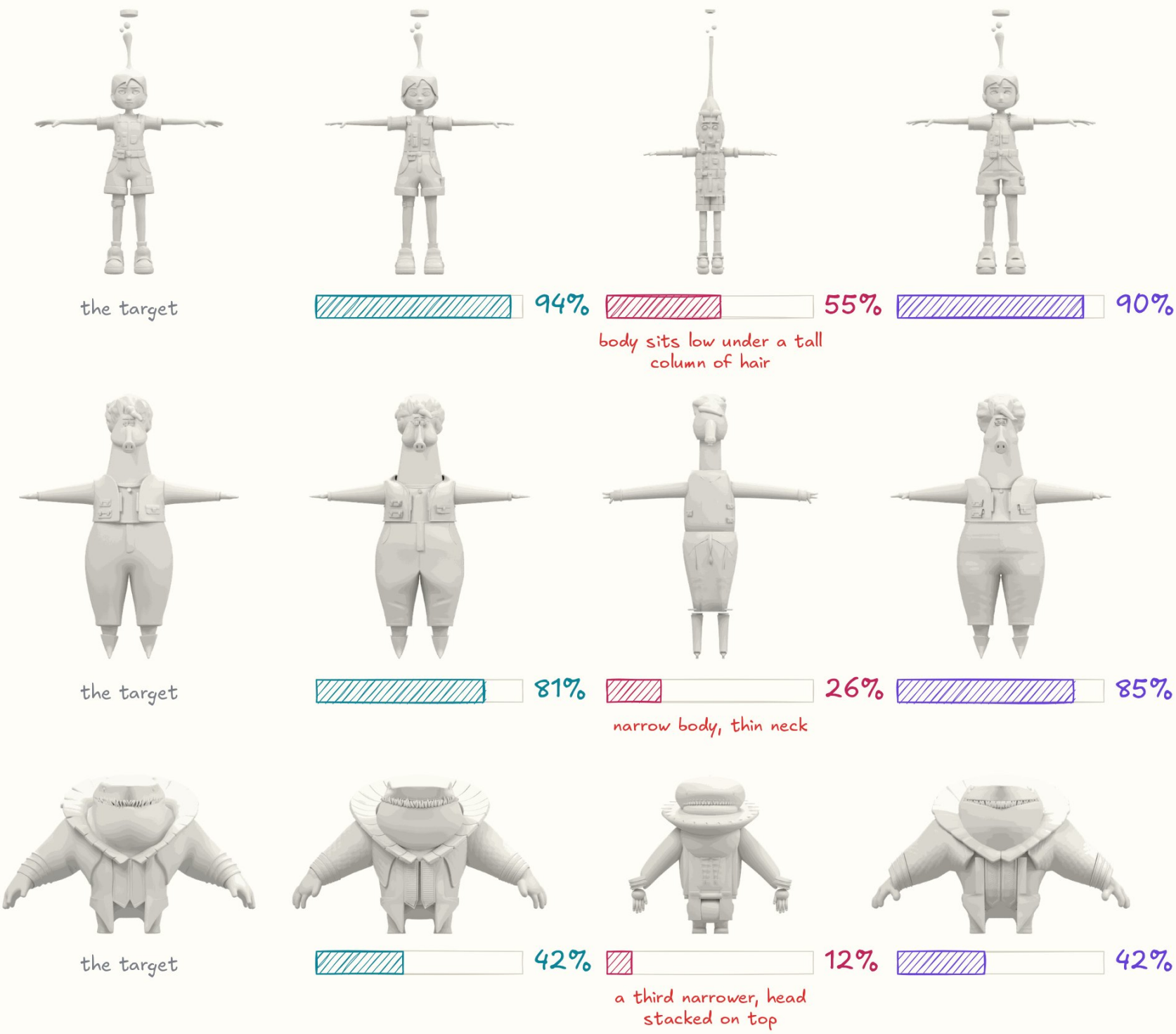


02 How close they got

Two models get the shape close. Brick is hard for all three.

We scaled each build to the height of the artist's model, stood both on the same floor and compared the two surfaces with an F-score, a standard measure in 3D reconstruction. It counts how much of each surface lies within a set distance of the other. We set that distance at 1% of the character's height, about 2 cm on a person-sized figure. A perfect match scores 100%.

Astra and Argon score between 81% and 94% on Lumi and Bess, and their builds match the artist's width and depth to within about 2%. Muse Wonder changes the proportions in every build, and it finished each run in under an hour, while Astra and Argon worked for 84 to 196 minutes. On Brick, the closest builds reach 42%.



Bars: F-score at 1% of the character's height. Figures: the saved geometry of every build, rendered in the same gray clay with the same camera, scale and light.

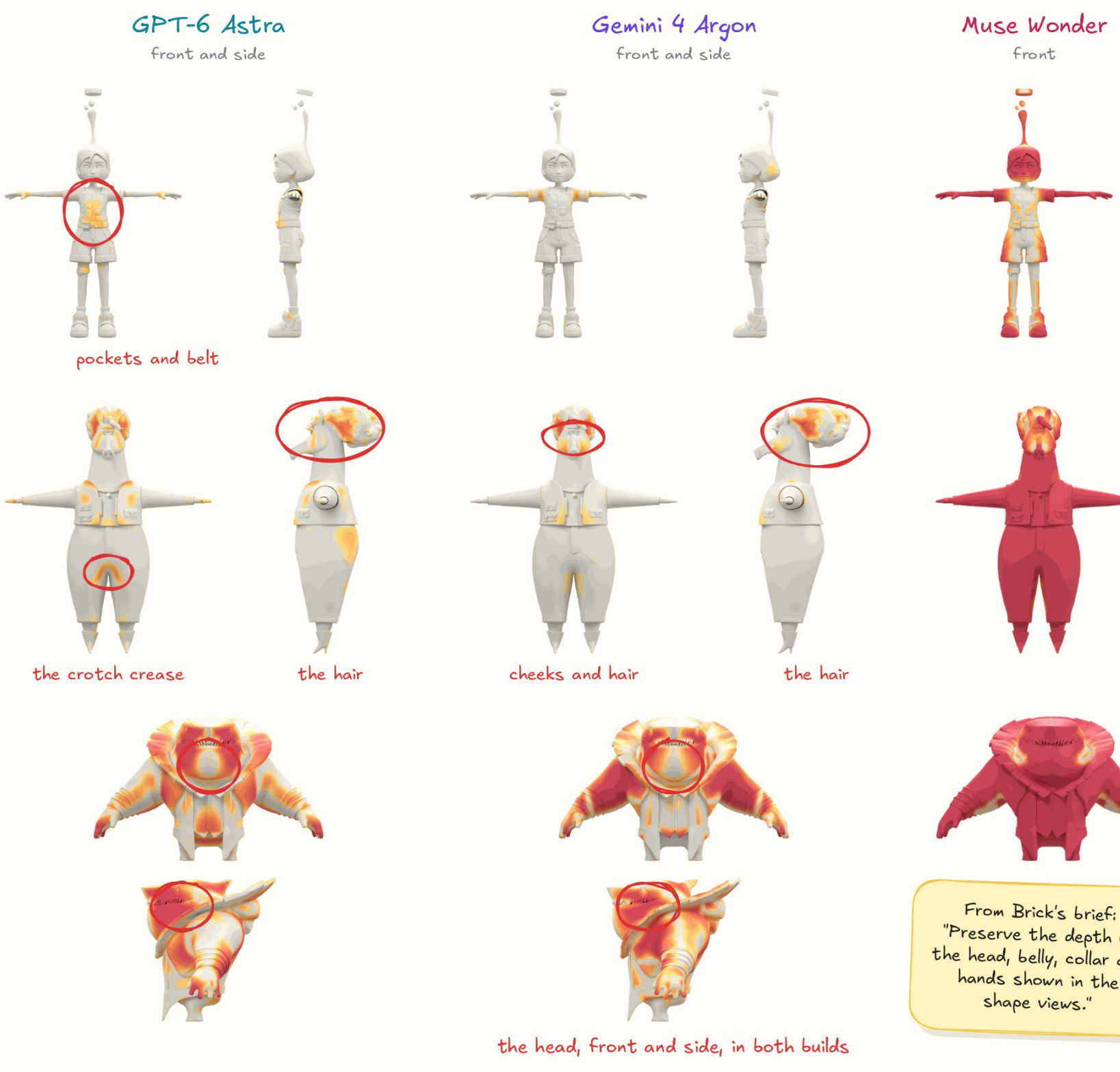
03 Where they miss

Small misses on Lumi and Bess. A large one on Brick's head.

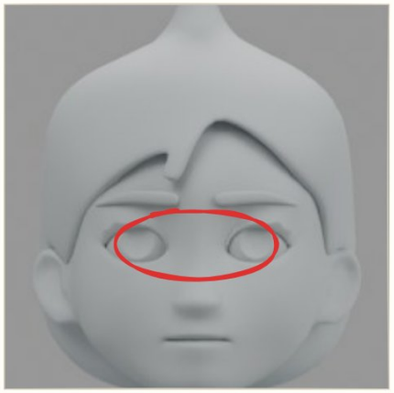
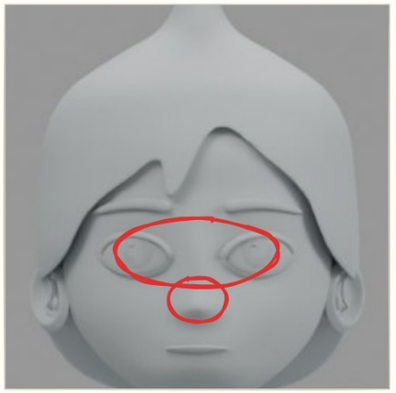
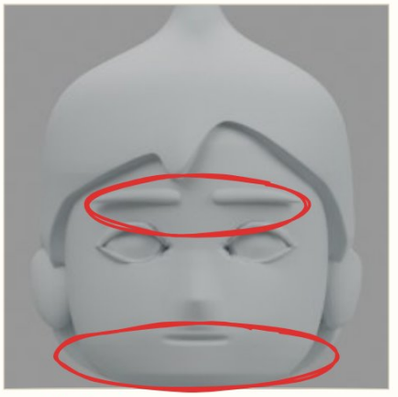
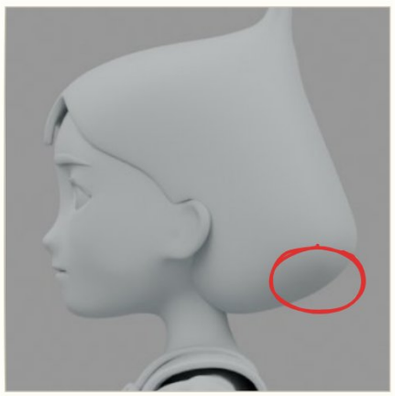
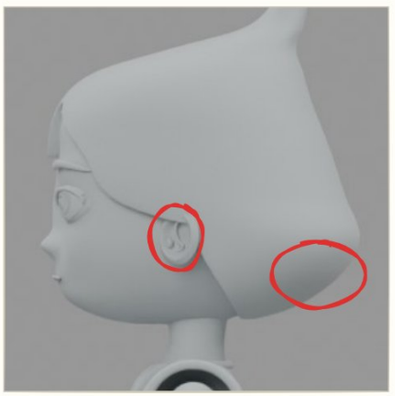
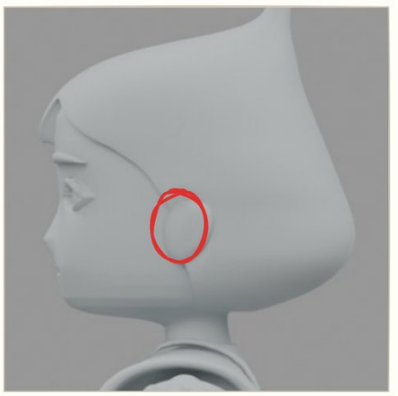
These maps paint the artist's model by how far each build's surface is from it. Gray means within half a percent of the character's height. Red means 4% or more.

For Lumi and Bess, the Astra and Argon maps are mostly gray. The warm patches sit on Bess's hair and cheeks, on pockets and belts, and on the hands. On Brick, more than half of the head's surface is over 2% of the height away in both builds. Muse Wonder's maps are red over most of the body.

distance from the artist's surface, as a share of height: within 0.5% 1% 2% 4% or more



Lumi: the right outfit, a different face

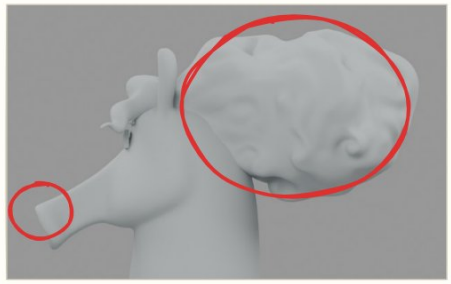
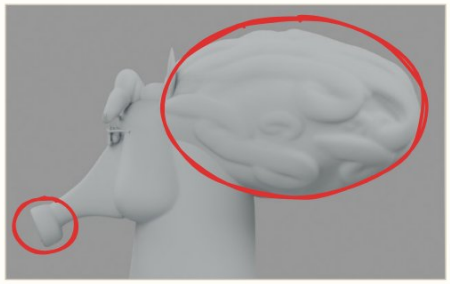
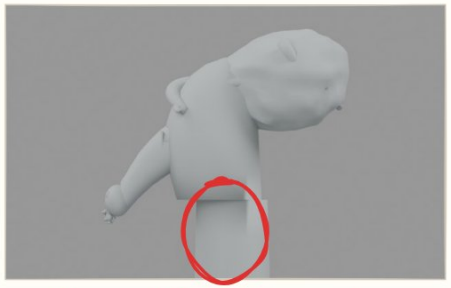
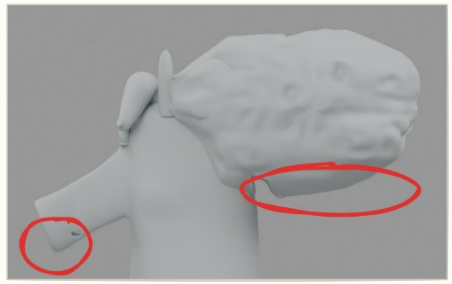
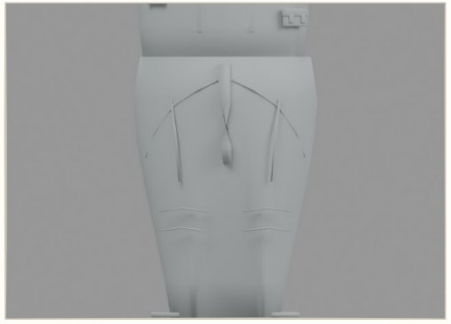
Artist	GPT-6 Astra	Muse Wonder	Gemini 4 Argon
 heavy lids, slight frown: serious	 thin even rims, soft nose: wide-eyed	 a face built from pieces	 blocky brows, square jaw: stern
 the bob curves under	 a hard edge at the back, a detailed ear	 eyes and cheeks stand off the head	 flat disc ears

Astra's Lumi is the closest build in the study. Its pockets, belt and rolled hems sit where the artist put them. The face is where it departs. The artist's heavy lids and slight frown make Lumi look serious. Astra's thin, even rims and softer nose make her look wide-eyed, and Argon's blocky brows and square jaw make her look stern.

In profile, the artist's bob curves under at the back, and Astra's ends in a hard edge. None of these changes is large. In both closer builds, most of Lumi's face is less than 1% of her height away from the artist's surface.

Lumi's face is about 3% of her surface, so these changes barely move a whole-body score.

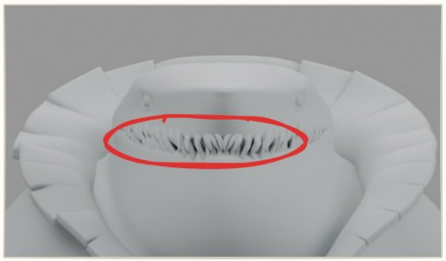
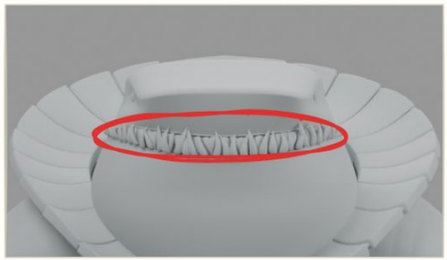
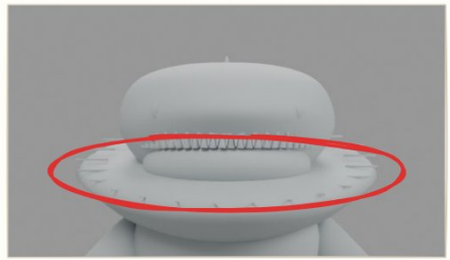
Bess: the right silhouette, a different weight

Artist	GPT-6 Astra	Muse Wonder	Gemini 4 Argon
			
muzzle ends flat and square; compact, lumpy bun	muzzle rounds down; long bun, combed ridges	the neck becomes a column	muzzle juts further; bun flat underneath
			
one smooth sweep	a sharp crease down the middle	narrow, straight legs	bands across the thighs

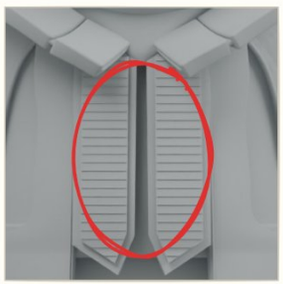
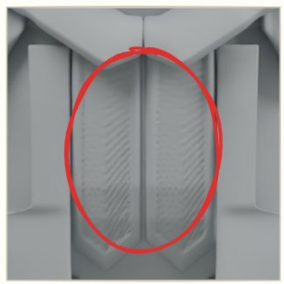
Astra and Argon keep what makes Bess recognizable: the long neck, the low glasses, the swept forelock. The differences show in profile, in the tip of the muzzle and the shape of the bun. Every bun still reads as stylized hair, and each puts a different weight on that long neck.

The trousers show a habit we saw in several builds: detail the artist did not put there. Astra adds a sharp crease down the middle and Argon adds bands across the thighs. The artist's trousers are one smooth sweep from waist to knee.

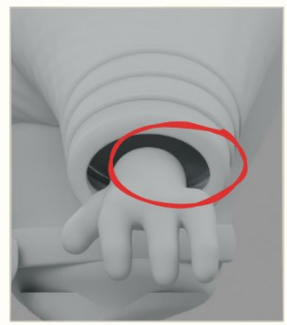
Brick: the same score, different problems

Artist	GPT-6 Astra	Muse Wonder	Gemini 4 Argon
			
uneven teeth that follow the curve of the jaw	a thick, even band	a small round head on a flat ring	a narrow slit

Chest panels

Artist	GPT-6 Astra	Gemini 4 Argon
		
plain	tight horizontal ribs	a fine diagonal pattern

Wrists

Artist	GPT-6 Astra	Gemini 4 Argon
		
the hand flows out of the cuff	tucked into the sleeve	a dark gap: the cuff looks like an empty tube

Astra and Argon both get Brick's composition: a wide head inside a collar that dips to a point in front and rises behind the shoulders. Muse Wonder sets a small, round head on a flat ring.

Up close, the two closer builds go wrong in different places: the mouth, the chest panels and the wrists.

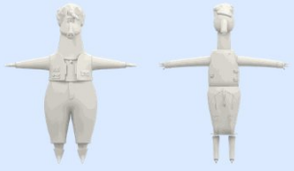
Astra and Argon both score 42% on Brick. An artist would send them back with different notes.

Two kinds of miss

The differences we found fall into two groups.

Shape

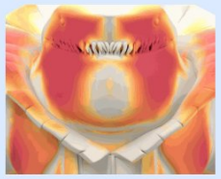
Large enough for a score to catch



artist Muse Wonder

Proportions

All three Muse Wonder builds.



Astra, error map

Large volumes

Brick's head, in every build.

The scores show these: Muse Wonder's builds score 12% to 55%, and no Brick build passes 42%.

Character

Too small for a whole-body score to see



artist Astra

Expression

Eyelids, a nose, the tip of a muzzle, a row of teeth.



Argon

Joins

Hair and jaw, sleeve and wrist, head and collar.



Astra

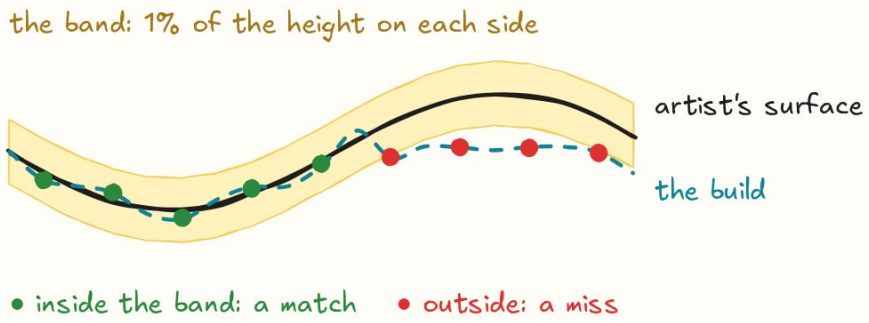
Detail that doesn't belong

Ribs, creases and bands where the artist left a surface plain.

The scores miss these: Astra's Lumi scores 94% with a different face.

For Astra and Argon, the second group is most of what an artist would ask them to change.

How to grade a character model



Geometry scores make a good first filter. Chamfer distance measures the average gap between two surfaces. F-score measures how much of each surface falls within a tolerance of the other. Both are standard in 3D reconstruction research, and both are cheap enough to run on every build. They catch wrong proportions and missing volume.

They also weigh every square centimeter the same, so a pocket counts as much as an eyelid. Artists review a character differently.

What artists look at

- ☑ Silhouette and proportions, from every side
- ☑ The volume of the main forms
- ☑ The face, and the expression it makes
- ☑ How parts join: hair to head, sleeve to wrist, head to collar
- ☑ How the clothes are built: thickness, layers, folds, and detail that belongs
- ☑ A clean, editable file, with topology that will deform well if the character is animated

A useful evaluation needs both: scores to find the large errors, and artists working from close views to find the rest.

09 Next

What's next

Nine builds are enough to see where the models struggle, but not enough to rank them. The next step is a blind review by character artists, using the checklist above, and a repair test: can a model fix a build when it gets an artist's notes?

Blind reviews by character artists

A repair test: from an artist's notes to a fixed build

An evaluation and a training set built from artists' corrections

If you build models for 3D work, we'd like to hear from you.

Notes on method

Models

GPT-6 Astra (reasoning effort max), Muse Wonder 1.0 (max) and Gemini 4 Argon (high, its highest setting), each through its provider's API with default sampling. All three had the same tools: control of a Linux desktop, Blender 5.2 through its interface and Python, a shell, a browser and web search. Each run started in a fresh workspace with an eight-hour limit. One Argon run on Lumi stopped on an error in our test harness and was rerun with the same inputs.

Inputs

A color render of the artist's model, gray renders of its shape from several angles with closer views of the head, hands or collar, and selected design art. Each brief also named details to get right, such as the depth of Brick's head and collar. Rigging and animation were not required.

All nine builds

		F-score at 0.5%	at 1%	at 2%	Chamfer, % of height
Lumi	GPT-6 Astra	80%	94%	99%	0.32
	Muse Wonder	37%	55%	71%	1.80
	Gemini 4 Argon	75%	90%	98%	0.38
Bess	GPT-6 Astra	64%	81%	92%	0.68
	Muse Wonder	16%	26%	41%	3.09
	Gemini 4 Argon	73%	85%	91%	0.64
Brick	GPT-6 Astra	26%	42%	63%	2.32
	Muse Wonder	6%	12%	20%	7.58
	Gemini 4 Argon	25%	42%	65%	2.33

Comparison

Every mesh is scaled to unit height, set on the floor and centered on its bounding box, with its pose unchanged. Lumi's transparent lamp is left out, as it is in her gray references. We sample 30,000 points per surface, weighted by area, and measure each point's distance to the nearest triangle of the other mesh. Chamfer distance averages the two directions. F-score is the harmonic mean of precision (the share of the build's surface within the tolerance of the artist's) and recall (the reverse), as in Tanks and Temples. Values are averaged over three samplings. Refitting each pair with a best-fit translation moves the 1% F-score by one point or less, except Muse Wonder on Brick (12% to 15%).

Pictures

Close-ups and lineups render the saved geometry in one gray material, with the same lighting and camera for all four models. Error maps color each vertex of the artist's model by its distance to the nearest surface of the build.

Sources

Knapitsch et al., Tanks and Temples (2017): precision, recall and F-score for 3D surfaces

Mescheder et al., Occupancy Networks (2019): Chamfer distance and normal consistency, reference code