research · 2026-08-24

What language are agent skills written in?

Plicara Research · 2026-08-24 ·

In the first quarter of 2026, 13.0% of newly written agent skills were in a language other than English, and one quarter later it was 16.3%. That is three points in three months across 255,068 skills, with confidence intervals nowhere near touching. For comparison, GitHub-wide non-English documentation took ten years to travel from 3.7% to 13.0%, so whatever is happening here is happening at a different speed entirely, and the most plausible explanation is that AI development has arrived somewhere other than San Francisco.

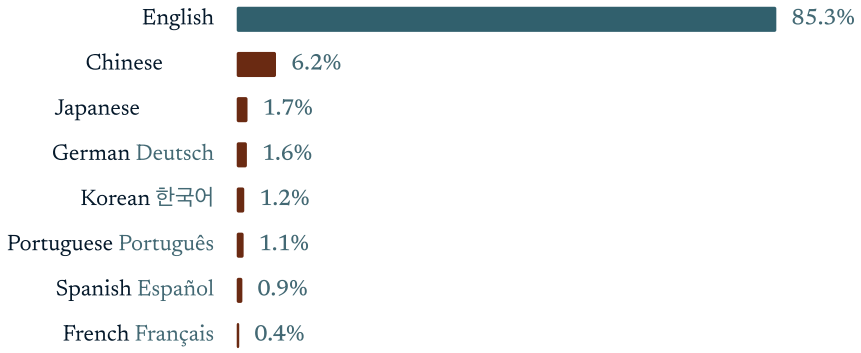
Reviewing the data, it turns out that the claim is stronger than the obvious version of it, because **English is not a proxy for American**. GitHub's fastest-growing developer population by a wide margin is India, which writes in English, as do Nigeria and Singapore, so a language count cannot see any of them. The non-English share is therefore not a measure of how much of this ecosystem sits outside the United States. It is a floor beneath it, and everything below should be read that way.

For context, a skill is a `SKILL.md` file in a folder, holding instructions for an AI agent in plain prose, loaded when the agent judges the task relevant. Anthropic published the specification in October 2025, and it spreads the way a recipe spreads: somebody copies it. Nine months later there were 3.8 million of them across 282,200 public repositories, which is what the GitSkills dataset collects. Skills are strange as software, by which we mean the traditional kind, because this is one of the things AI has upended. They are written in human language and the runtime is a multilingual model, so there is no technical reason to write one in English: a developer in Shenzhen or São Paulo can state a procedure more precisely in their own language, and the agent will follow it. Whether it follows it as *well* is a better question, and much harder to answer than anything a file crawl can settle.

the distribution

We ran language identification over the prose body of every distinct skill, after stripping front matter and fenced code.

HORIZONTAL BAR CHART, LANGUAGE DISTRIBUTION



1,870,299 distinct skill contents. The 14.3% that are not English are led by Chinese.

LANGUAGE	SHARE OF DISTINCT SKILLS
English	85.3%
Chinese	6.2%
Japanese	1.7%
German	1.6%
Korean	1.2%
Portuguese	1.1%
Spanish	0.9%
French	0.4%

So 14.3% of skills are not in English, and split by script the Chinese ones run 104,985 simplified against 9,112 traditional. The rows above do not quite sum to that, because 6,810 skills came back below our confidence floor and are counted as neither. The comparison worth making is against GitHub's own documentation instead of its issues or pull requests, and a 2026 ICSE study put repository documentation at 13.0% non-English, with Chinese at 3.3% of repositories. In aggregate that makes skills unremarkable, 14.3% against 13.0% being a dead heat. They are markedly more Chinese, though, 6.2% against 3.3%.

why every published number disagrees

Ours is not the only published figure, and the published figures do not agree with each other.

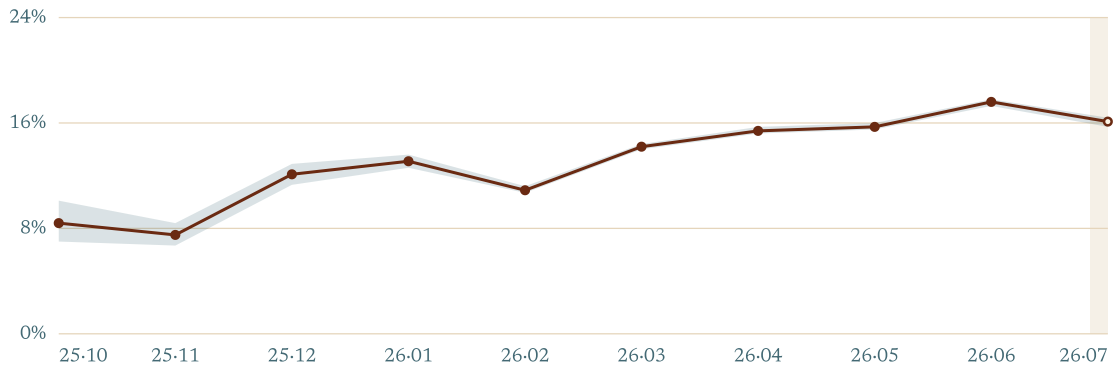
REPORTED ENGLISH SHARE	CORPUS	METHOD
65.0%	557 healthcare skills, ClawHub (2605.02709)	not stated
81.8%	26,502 skills, ClawHub (2604.13064)	not stated
85.3%	**1,870,299 distinct, GitHub (ours)**	py3langid, conf >= 0.80
92.6%	133,149 skills, skills.sh (2607.01456)	fast-langdetect
99.7%	English-seeded crawl (2606.03565)	seeded

These are not contradictions, they are five different populations: curated marketplaces skew English, domain slices skew toward wherever that domain happens to be active, and a crawl seeded with English queries will find English. The first candidate to rule out is us, because if our identifier simply saw less English than everyone else's then the whole comparison would be an artifact of tooling. So we ran both over the same documents, py3langid which we use and fast-langdetect which the 92.6% study used. They agree on 97.6% of documents, and their English shares sit +1.2 points apart against a gap of around seven. Quality screening looks like the next good candidate and leads nowhere either: if corpora that filter for valid front matter were quietly discarding non-English skills that would explain some of the spread, but non-English skills have slightly *better* front-matter validity, 88.1% against 86.6%, and filtering moves the English share only from 85.6% to 85.4%. What is left is where you looked. That generalises well past this dataset, so when someone tells you what "the AI ecosystem" looks like, the registry they scraped may hold more of the answer than anything else they say.

skills are getting less English

Skills carry commit history, so each one has a creation date, and that turns a static pie chart into a trend.

NON-ENGLISH SHARE BY MONTH, WITH CONFIDENCE BAND

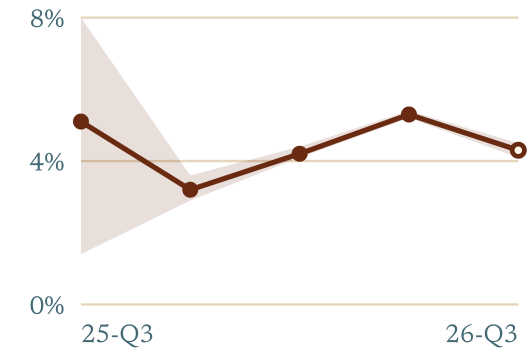


Band is the 95% Wilson interval. July 2026 is shaded: collection ran mid-month, so that cohort is censored and excluded from comparisons.

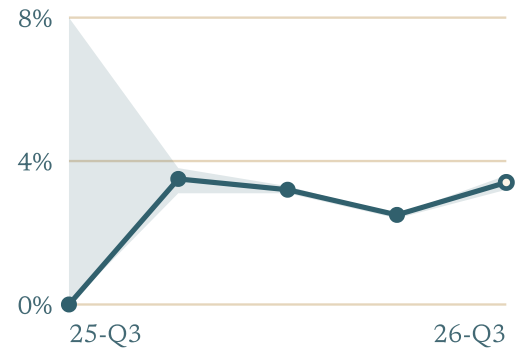
QUARTER	NON-ENGLISH SHARE
2026 Q1	13.0% [12.8, 13.1]
2026 Q2	**16.3%** [16.1, 16.4]

Month by month the climb is not smooth, since February dips to 10.9% before March resumes at 14.2%, but the direction across the window is not in doubt: 13.1% in January against 17.6% in June. That is roughly what you would expect of a format eighteen months old, since new artifact types acquire their demographics much faster than mature ones when there is no incumbency to overcome. But "non-English" is not one thing, and broken out, the rise turns out to be carried by two of the four groups rather than by all of them.

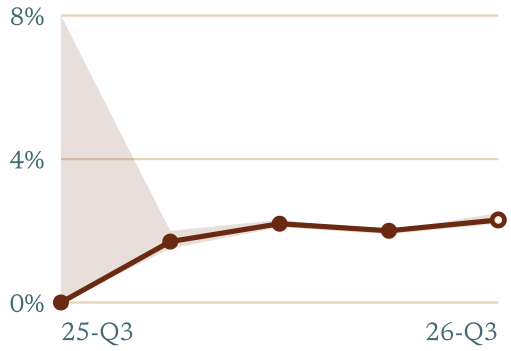
SMALL MULTIPLES, SHARE BY QUARTER PER LANGUAGE, WITH CONFIDENCE BANDS



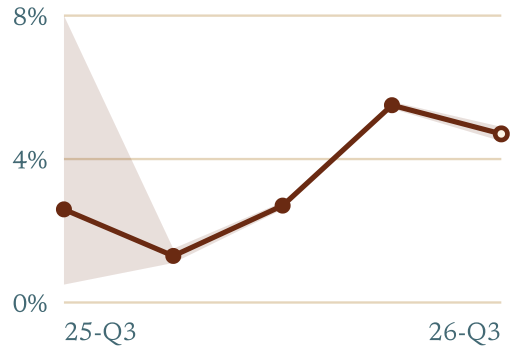
Chinese **+2.1 pts** 25-Q4→26-Q2



Japanese **-1.0 pts** 25-Q4→26-Q2



Korean +0.3 pts 25-Q4→26-Q2



European +4.2 pts 25-Q4→26-Q2

Shaded band is the 95% Wilson interval; the hollow final point is the censored July cohort, plotted but never compared. European groups German, French, Spanish, Portuguese, Italian, Russian and Dutch.

	2026 Q1	2026 Q2	CHANGE
Chinese	4.2% [4.1, 4.4]	5.3% [5.2, 5.4]	+1.1
European	2.7% [2.6, 2.8]	5.5% [5.4, 5.6]	+2.8
Korean	2.2% [2.1, 2.3]	2.0% [1.9, 2.0]	-0.2
Japanese	3.2% [3.1, 3.3]	2.5% [2.4, 2.5]	-0.7

European languages, by which we mean German, French, Spanish, Portuguese, Italian, Russian and Dutch grouped together, more than double across the window while Chinese climbs steadily, and Japanese and Korean do neither: Japanese was the most common non-English language at the end of 2025 and slipped through the first half of 2026 as everyone else arrived, while Korean stays flat throughout. The censored July cohort hints that Japanese is recovering, and we are not counting it. Changes are measured between the two complete quarters, 2026 Q1 and Q2, since the final column is the July collection month and is censored, so it appears in the chart but never in a comparison.

why we believe it

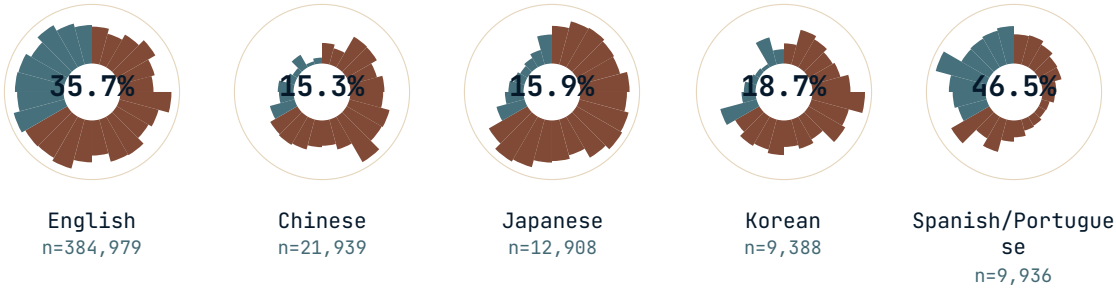
A trend like this is exactly the kind of thing that turns out to be an artifact, so we spent longer trying to break it than we did finding it. Commit history exists for only 24% of skills, and that subsample leans toward heavily copied ones, which matters because copying turns out to be strongly related to language. The worry, in other words, is that we are watching a

selection effect and not a change in what people write. Holding copies fixed at one, the rise is larger than the headline, 14.7% to 18.1%; counting each repository only once, so that no bulk uploader can swing it, the rise survives at 14.5% to 16.8%.

the clock

Commit timestamps are stored in UTC, so an author's local timezone is gone before we ever see the file. But people mostly commit while they are awake, and if a group of skills is written by people in one part of the world, their commits should vanish during that region's night.

24-HOUR DIALS, ONE PER LANGUAGE



Centre figure is the share of first commits in that window. The non-English groups are small, so read the contrast, not the decimals.

LANGUAGE	COMMITS DURING EAST ASIAN NIGHT	N
English	35.7%	384,979
Chinese	**15.3%**	21,939
Japanese	15.9%	12,908
Korean	18.7%	9,388
Spanish/Portuguese	**46.5%**	9,936

Chinese-language skills fall to less than half the English rate in that window, while English itself stays flat across all twenty-four hours, which is the signature of a globally distributed population with no single night. Spanish and Portuguese run the opposite way and peak at 19:00 UTC, mid-afternoon in Brazil and late evening in Iberia, which places those authors

in the Americas. Nothing in the language identifier knows what time a file was committed, so the two signals are independent, and they agree.

Honest limits. This is a population-level phase estimate, good to a couple of hours at best; it cannot separate UTC+8 from UTC+9, a language is not a country, and it says nothing whatsoever about any individual author. We found no published validation of hour-of-day inference at this granularity, so treat it as corroboration and not as geolocation. A raw git commit does record the author's UTC offset, and this dataset normalised it away, which is the fix for anyone building on this.

the same story from outside

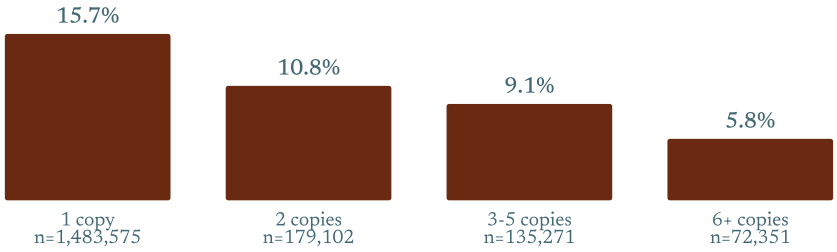
We are reading one artifact type on one platform, so the question that matters is whether anyone measuring something else sees the same movement, and they do, at a larger scale than we can.

GitHub's own Octoverse 2025 reports that India added 5.2 million developers in a single year, about 14% of the 36 million accounts opened worldwide, which takes it to 21.9 million and second place globally. That is 4.9 times its 2020 population. Brazil grew 4.1 times over the same period, Indonesia 4.8 and Japan tripled, and new signups now run at roughly 25 a minute across APAC against 12 across Europe. India, Brazil and Indonesia together account for about half of all new accounts. Stanford's 2026 AI Index puts generative-AI adoption at 64% in the United Arab Emirates and 61% in Singapore, against 28.3% in the United States, which ranks twenty-fourth. Its policy chapter records open-source contributions "from the rest of the world now outpacing Europe and approaching the United States on GitHub", and its education chapter finds AI engineering skills accelerating fastest in the UAE, Chile and South Africa.

None of that is about agent skills, which is exactly what makes it useful: three independent measurements of where AI development is happening, none of them looking at `SKILL.md` files, all of them pointing the same way. Our number is the same phenomenon surfacing in a corpus eighteen months old. Our number also cannot see most of it. India writes in English, so the largest single engine of GitHub's growth is invisible to a language count, and so are Nigeria and Singapore, which is why 14.3% should be read as the floor beneath whatever share of this ecosystem now sits outside the United States.

copied, or tended?

NON-ENGLISH SHARE BY COPY COUNT

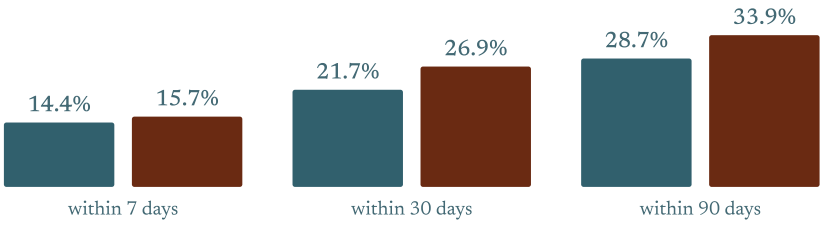


Across all 1,870,299 distinct contents. Removing the ten aggregator repositories, or counting distinct owners, widens the gap.

English skills get copied more. The non-English share falls steadily the more a skill is reused, from 15.7% among skills nobody has ever copied down to 5.8% among those copied six or more times. That one needed defending, because a great deal of what looks like copying on GitHub is really archiving. Ten repositories hold 14.5% of every skill file here, and they are registries and mirrors rather than authors, so 282,200 repositories behave, by concentration, like about 250. Those aggregators turn out to lean non-English, 20.5% against 13.7% elsewhere, so whatever they are doing to the numbers works against this pattern instead of producing it. Excluding them leaves the gap where it was, 15.1% down to 5.3%, and counting distinct owners, so that one actor vendoring a skill into ten of their own repositories counts once, widens it slightly to 15.1% against 4.6%.

But non-English skills get revised more. That needs an age correction, since non-English skills are younger on average and have had less time to be touched, so the comparison below holds age fixed and asks what share of skills at least N days old were revised within their first N days.

REVISION RATE AT 7, 30 AND 90 DAY WINDOWS



Compared at equal age: among skills at least N days old, the share revised within their first N days.

WINDOW	ENGLISH	NON-ENGLISH
7 days	14.4%	15.7%
30 days	21.7%	26.9%
90 days	28.7%	33.9%

The gap opens over the first month and then holds, +1.3 points at a week, +5.2 at a month and +5.2 at three, which leaves two populations behaving quite differently: English skills propagate, written once and copied widely and rarely touched again, while non-English skills are tended, copied less and revised more. The likely reason for the copying half is search. Discovery is lexical, and a developer searching in English will not surface a skill written in Chinese even where a multilingual model could execute it perfectly, so what stands between a Shenzhen developer's skill and the person who needs it is a text match. If that is right, the ecosystem globalises in what gets written well before it globalises in what gets reused, and the lag between the two is a tooling problem somebody could fix.

who is actually writing these?

The GitSkills authors asked one more thing worth answering: how many skills do agents write themselves? The obvious way to check fails immediately, because GitHub's own bot flag catches almost nothing: agent-written code is committed under the human's account. The signal that does work is the trailer, the `Co-Authored-By` line a coding agent appends to commits it authored, which is the tool claiming authorship instead of us inferring it from prose.

MEASURE	VALUE
Name an AI agent in a commit trailer	**30.4%** [30.3, 30.6]
Flagged as a bot by the platform	1.0%
Japanese skills, agent-authored	43.4%
Chinese skills, agent-authored	23.2%

Nearly a third of skills carry an agent's fingerprint and the platform sees almost none of it, with Claude accounting for the overwhelming majority of those trailers while Cursor,

Copilot and Codex trail far behind, and the trailers even carry model versions. Read it as a floor, since a skill whose author stripped the trailer, or squashed it away, or used a tool that never emits one, counts as human here.

what this does not show

Language identification keys on script and function words, so a German skill thick with English technical vocabulary gets pulled toward English, which is another reason the non-English share is a lower bound. Dates cover only part of the corpus and only deduplication representatives, so "created" means the first commit touching *that copy* and not the first appearance of that content anywhere. And a crawl sees only survivors, so any skill created and deleted before July 2026 is invisible to us, which inflates every maintenance figure here by an amount we cannot estimate.

what we would want to know next

The number we cannot get from this data is whether any of it costs anything. A skill written in Chinese and never copied might be worse, or might be identical work that nobody found, and those two worlds look the same from a file crawl while implying opposite things. Separating them needs execution traces, and if somebody has those we would like to see them. The larger question is what the floor actually rests on. 14.3% of skills are not in English and the share is climbing three points a quarter, while the fastest-growing developer population on the platform writes in English and never appears in that count at all. Whatever the real number is, everything we can measure says it is moving in one direction, and faster than anything comparable has moved before.

Article 02 stays on this corpus and asks which *programming* languages skills talk about. Our first pass had Shell/Bash leading every language at 37.5%, which turned out to be an artifact of counting pasted commands, and measured by what a skill actually ships Python leads at 7.6% while Shell drops to 3.1%.

credit

None of this exists without the dataset, which was built and released by someone else, and all credit for collecting, deduplicating and documenting 3.8 million skill files belongs to its authors:

> Giuseppe Destefanis, Daniel Graziotin, Matteo Vaccargiu, and Marco Ortu. 2027. GitSkills: A Dataset of Agent Skills on GitHub. In *Proceedings of the 24th International Conference on Mining Software Repositories (MSR '27).*

Preprint arXiv:2608.10906, archive 10.5281/zenodo.21875637, Parquet mirror `mvaccargiu/gitskills`, sample `giuseppedestefanis/gitskills-sample`, licence CC BY 4.0.

GitSkills is the dataset for the MSR '27 Mining Challenge, and we are not affiliated with its authors, with MSR, or with the challenge, so nothing here should be read as endorsed by them: the dataset is theirs, and the analysis and any error in it is ours. Several of the research questions we take up, including which natural languages skills are written in, are ones the GitSkills authors posed and left open.

Analysis code lives at github.com/plicara/articles under `gitskills-analysis/`, where every figure is generated from a single machine-readable export and never typed by hand, so the whole thing can be regenerated and checked.

Found something wrong? We would genuinely like to know.