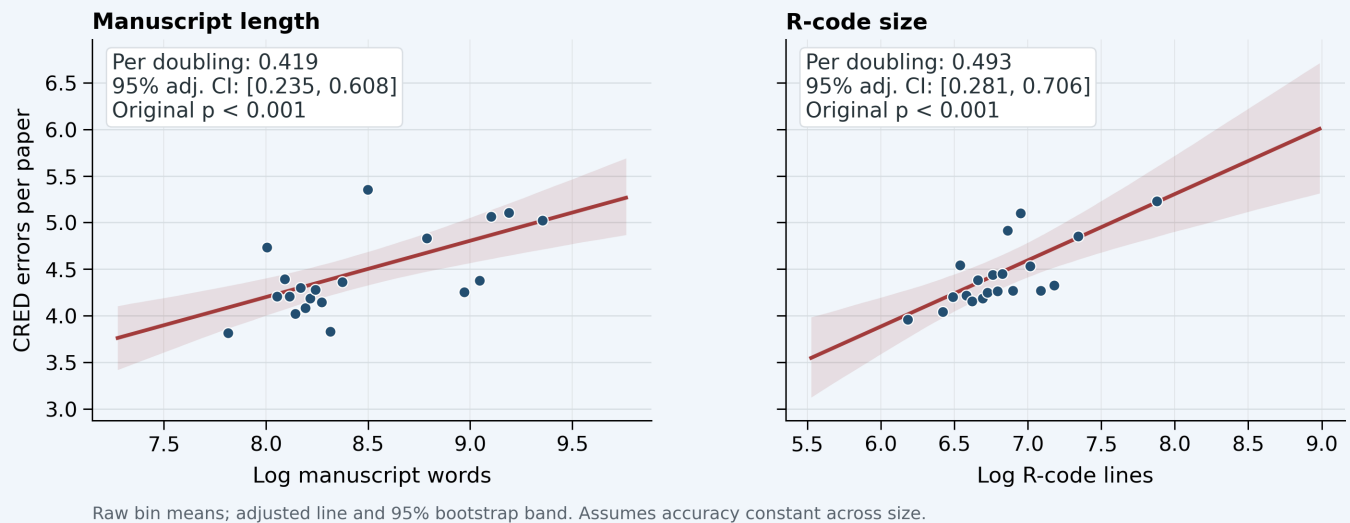


Does paper complexity increase errors?

We investigate whether more complex papers contain more errors. Across Project APE's first 1,000 papers, larger manuscripts and code bundles attract more reported errors, especially paper-code mismatches. Prose-table mismatches move the other way. The distinction matters for how we assess research reliability.

1 | Total CRED errors

N = 957



The hypothesis

A paper's claims, tables and code must agree. Our hypothesis is that this becomes harder as research grows more complex, leading to more errors. We examine that prediction in [Project APE's first 1,000 papers and their code](#).

We cannot observe complexity directly, so we use two proxies: manuscript words, excluding LaTeX commands and other non-prose material, and lines of R code, excluding blank lines and full-line comments. Both measure size; neither captures how difficult an argument or a computation is.

Our sample is the 957 papers with R code and positive counts on both measures; the other 43 have no R source. Tournament ratings exist for 949 of the 957.

What the counts show

On average, our Codex [CRED verifier](#) flags 4.43 problems per paper. That count rises with size on both measures. Doubling a manuscript's words is associated with 0.42 more errors; doubling its R code, with 0.49 more. These estimates allow for verifier accuracy; each is roughly a tenth of the observed average.

The dots average 20 groups of similarly sized papers. In Figures 1-3, lines adjust for verifier accuracy; shading shows 95% bootstrap bands, including uncertainty in the relationship and validation rates.

These are two separate comparisons. Neither holds the other measure constant, and neither adjusts for topic, writing model, date or other differences between papers.

What this says about reliability

Our Codex 5.5 high verifier has strong [validation evidence](#). Human reviewers confirmed 97.4% of its findings; single-error benchmark recall was 96.5%.

At those rates, correcting for missed and invalid findings barely changes Figure 1's slopes. Both stayed positive in 20,000 simulations assuming accuracy constant across paper size. Size-dependent accuracy remains a limitation.

More prose and code also create more opportunities for genuine errors. Higher counts need not mean that each claim or operation is less reliable, or that complexity causes errors. To understand which problems drive the pattern, we examine paper-code and prose-table mismatches.

What we find by error category

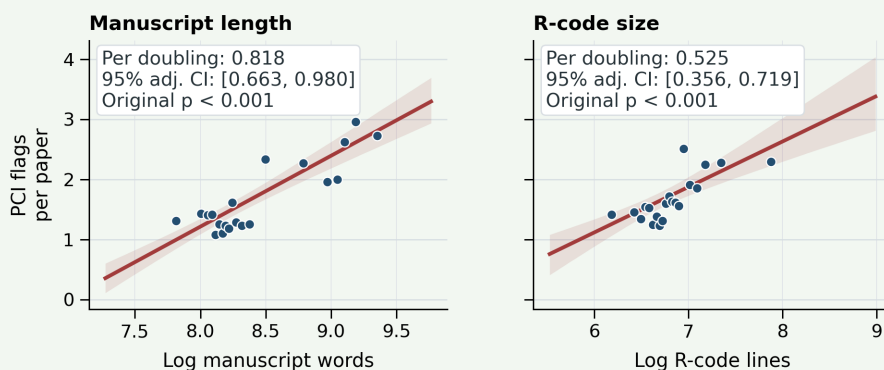
Where paper and code diverge

The clearest increase is in paper-code inconsistencies (PCI), where the manuscript says one thing and the code does another. Doubling manuscript words is associated with 0.82 more such errors; doubling R code, with 0.52 more. Both slopes remain positive after allowing for PCI precision (97.7%) and recall (94.4%).

Larger papers warrant closer checks of written methods and results against code. This association does not establish that complexity caused the mismatches.

2 | Paper-code mismatches (PCI)

N = 957



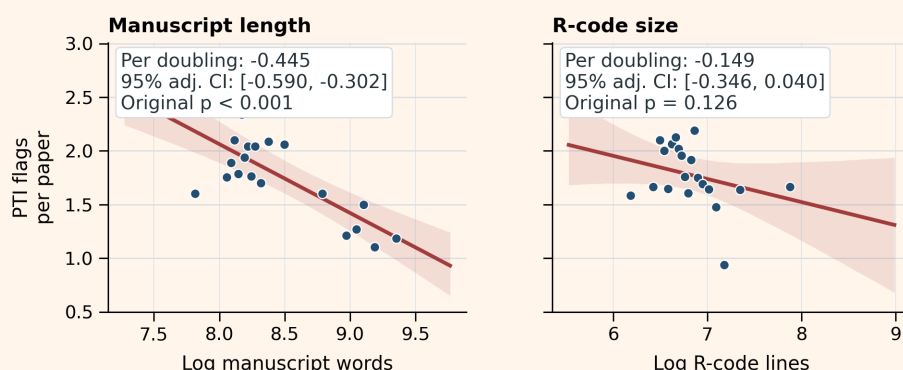
Prose-table mismatches decline

The pattern reverses for prose-table inconsistencies (PTI), where the text and the tables it describes disagree. Doubling manuscript words is associated with 0.44 fewer such errors. For R code the decline is smaller and statistically uncertain: 0.15 fewer per doubling ($p = 0.126$).

The adjustment uses PTI precision of 96.3% and recall of 96.0%. The contrast with PCI remains: different types of error need not accumulate together as papers grow.

3 | Prose-table mismatches (PTI)

N = 957



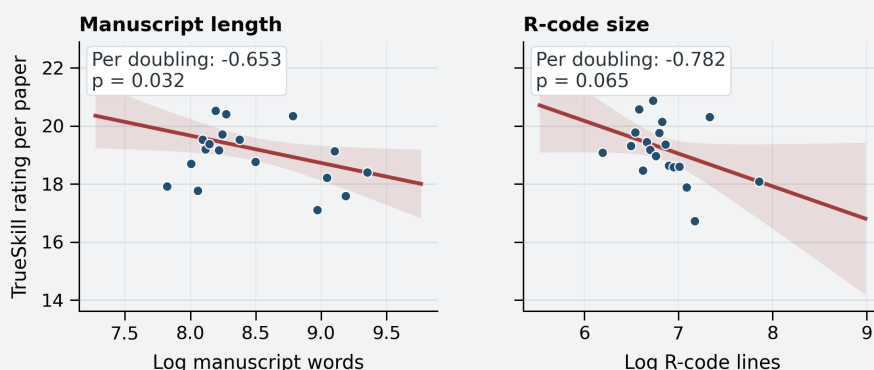
Do judges reward complexity?

We also ask whether larger papers earn stronger evaluations, using [TrueSkill ratings](#) from our [LLM-judge tournament](#), whose judge reads only PDFs, never code or data, and whose ratings depend on opponents and match history.

We find no clear reward for greater size. Doubling manuscript words is associated with a rating 0.65 points lower ($p = 0.032$); doubling R code, with one 0.78 points lower ($p = 0.065$). These are modest signals from a PDF-only judge, not a case for shortening papers.

4 | Scientific value: LLM-judge rating

N = 949



Method and limits. We fit separate bivariate regressions on natural log size, excluding papers with zero size measures. Figures 1-3 use 10,000 joint paper/validation bootstrap draws and chapter-specific accuracy for PCI/PTI; Figure 4 uses HC1 bands. Calibration assumes accuracy transfers across sizes; it does not correct chapter misclassification. P-values test original slopes, without adjustment for eight comparisons. Repeated manuscripts and shared tournament histories may affect uncertainty. Ratings predate the CRED scans. These associations do not establish causation or measure error rates per claim.

Sources. Papers and code: [Project APE's public repository](#). Error definitions and verifier validation: [Verifying the Verifiers](#), using CRED v0.92 and Codex 5.5 high. Scientific-value ratings: [Project APE's tournament methodology](#).