

Erratum to: The Alignment Squeeze: Modelling Competing Correction Pressures in Language Models

Nell Watson¹ and Rich Dalton²

¹ University of Gloucestershire, School of Computing and Engineering,
The Park, Cheltenham GL50 2RH, United Kingdom
`eleanorwatson@connect.glos.ac.uk`

² Independent Researcher, Independence, Missouri, United States
`radalton30@gmail.com`

Erratum dated 6 October 2026. The paper describes its twelve-prompt *focused set* as the pilot prompts that “showed neither ceiling nor floor compliance”, and says the focused-set estimate is inflated by “selection on variance”. Both descriptions are wrong: the released pilot and focused-set records show that the twelve are exactly the pilot prompts with a nonzero monitored–unmonitored difference in the three-sample pilot. One of them had no within-cell variance, and two prompts with variance but no difference were excluded. Selection was therefore on the effect itself, which inflates the twelve-prompt estimate more than selection on variance would, and the Llama-3.1-70B comparison ran on prompts chosen using the Qwen pilot, a further confound in the model $\times$ condition contrast.

No reported number changes. The paper’s estimate of the effect’s size rests on the 48-prompt factorial follow-up (three to six percentage points), which is unaffected. The records and corrected copies of both versions of the paper are at <https://quasiqualia.com/#alignment-squeeze>. The corrected sentences:

Location	Printed	Corrected
Sect. 3, scenario sets	the pilot prompts that showed neither ceiling nor floor compliance in a three-sample pilot	the pilot prompts that showed a nonzero monitored–unmonitored difference in a three-sample pilot
Sect. 3.2, Results	the focused-set estimate is inflated by selection on variance	the focused-set estimate is inflated by selection on that difference
Sect. 3.2, Cross-model	is confounded by quantisation	is confounded by quantisation and by prompts selected on Qwen’s pilot
Sect. 4, Limitations	Twelve prompts selected on pilot variance inflate the within-prompt estimate	Twelve prompts selected on their pilot difference inflate the within-prompt estimate