Ravenscar AI

RESEARCH PAPER / NZBENCH / NOVEMBER 2026 EDITION

Does AI get New Zealand right?

A benchmark of 1,000 everyday questions about New Zealand's rules and services, asked of eight AI models with and without a location.

Barnaby Carter-Scofield

Ravenscar AI, Christchurch, New Zealand • November 2026 edition • study specification 2.3

63%

of answers were complete when the question named New Zealand

22%

were complete when it did not

13%

of answers with no location were written for another country

Abstract

People increasingly ask artificial intelligence (AI) models everyday questions about wages, tenancy, health, benefits and immigration. Answers that are right for other countries can often be wrong for New Zealand. Ravenscar AI presents NZBench, a benchmark of 1,000 everyday questions initially covering 32 topics, each with an official answer quoted from its source. We asked eight widely used AI models every question twice, once as a person would type it and once beginning "In New Zealand:". Each answer was graded by an AI judge. Two checks stand behind the judge: a fixed script with no AI in it, for the 508 questions whose answer is a number, date or amount, and two people who graded a sample of the same answers. When the question named New Zealand, models gave a complete answer 63% of the time (38% to 79% across models). When it did not, they gave a complete answer only 22% of the time. When the question did not have a location, 13% of answers were written for another country, most often the United States, and 53% hedged across several. Models were weakest on figures that change every year: only 28% of answers gave the current figure, compared with 81% for rules that rarely change. The judge agreed with people on 84.6% of checked answers (82.2% weighted), below the 85% mark set in advance; it was stricter than people, never more lenient. The main findings hold on the script, which uses no AI. We release the public questions, answers, grades and code.

1 Introduction

People will only trust AI systems if those systems give them information that is correct. That means being correct and knowing where they live. New Zealanders increasingly put everyday questions about wages, tenancy, health and benefits to AI chatbots. In a 2025 survey of 1,003 online adults, 79% had used AI in the past year, and 61% were very or extremely concerned that it may produce inaccurate information (InternetNZ, n.d.). An answer that is right for another country can be wrong here. When people get what they perceive to be poor or inaccurate answers they may decide the technology cannot be trusted. Most users are not told which model they are using, or that the way a question is worded changes the answer.

New Zealand is a hard case for an AI model, for three reasons. It makes up a small share of the text that models learn from. Its rules are close to those of Australia and the United Kingdom, and its English is close to that of the United States, so one country's answer can pass for another's. And many of its figures change on a set date each year, so an answer can be for the right country and still be out of date.

AI safety and alignment is an increasingly important area of research, and in the author's view New Zealand is falling behind in it. Ravenscar AI built NZBench to help close part of that gap.

This paper asks three questions. How often do widely used AI models give a complete and correct answer to an everyday question about New Zealand? How much does the answer depend on the person saying where they are? And how far can an AI judge be trusted to grade the answers? We make four

contributions:

- **A benchmark.** 1,000 questions in 32 topics. Each has an official answer and the passage it is quoted from. We hold back 76 questions so that later AI systems can be tested on questions that have never been published.
- **A controlled test of location.** We ask every question with and without the words “In New Zealand:”. This shows how much a model’s answer depends on being told the country to give an accurate answer that a user can trust.
- **Three checks on the same answers.** An AI judge, a fixed script with no AI in it, and people who checked the judge. We fixed the design, the tools and the pass mark, and recorded a fingerprint of each, before any data existed (Nosek et al., 2018).
- **A reported failure.** The judge missed its pass mark. We report by how much, in which direction, and what a reading of every contested answer found.

2 Related work

Local knowledge and location. Benchmarks of local knowledge find that AI models do worse on local questions than on global ones (Etxaniz et al., 2024; Lee et al., 2024; Libovický et al., 2025; Myung et al., 2024). Most of them cover places where the local language is not English, so it is hard to separate the effect of language from the effect of place. New Zealand is an English speaking case. When a question names no country, a model has to guess, hedge or ask. Wang and Suresh (2026) found that 74.5% of responses to English legal and administrative questions used a United States framework when a single answer was required. Geographic bias has also been shown in models’ judgements more broadly (Manvi et al., 2024).

Out of date answers. Models are trained up to a fixed date in time, and the world moves on (Lazaridou et al., 2021). Without a way to look up current information, they give out of date information (Kasai et al., 2022; Vu et al., 2024). In law, they often apply a rule that is no longer in force (Prior et al., 2026). Studies of legal questions report high error rates for general purpose models (Dahl et al., 2024) and substantial error rates for legal research tools (Magesh et al., 2025).

AI judges. Grading open ended answers at scale usually means using an AI model as the judge (Zheng et al., 2023). A judge can favour text written by its own model (Panickssery et al., 2024). How often judges agree with people varies widely by task (Bavaresco et al., 2025). The standing advice is therefore to check a judge against human labels before trusting it (Bavaresco et al., 2025; Shankar et al., 2024). We follow that advice. We also report ranges that treat the question as the unit that varies (Miller, 2024).

3 The benchmark

Questions. NZBench has 1,000 questions in 32 topics that New Zealanders commonly need help with, from the minimum wage and tenancy to ACC, KiwiSaver, benefits and immigration. Each question has three parts: an official answer taken from an official or authoritative source, the passage the answer is quoted from, and a short list of the points a right answer must include. The questions and answers were drafted from the official pages with the help of an AI model, under the author’s direction. Drafting followed three rules. Nothing goes in an official answer that its source does not state. A word for word

quote is recorded for every question. Any question that could not be confirmed on a readable page is left out. Drafting favoured facts that had changed recently, that differ from other countries' rules, or that people commonly get wrong. The question bank is therefore harder by design than a random sample of everyday questions.

Two wordings. We ask each question twice: once as a person would type it, with no location, and once beginning "In New Zealand:". The difference between the two results is the location effect.

Three measures. *Complete*: the AI judge finds every required point in the answer, stated for New Zealand. *Key figure*: for the 508 questions whose answer is a number, date or amount, a fixed script finds New Zealand's figure in the answer. The script uses no AI. *Country*: which country the answer is written for. Complete is the strictest way to read an answer and key figure is the most lenient. A finding that holds on both does not depend on how strictly answers are read.

4 Experimental setup

Models. We asked eight AI models from five companies every question in both wordings on 14 and 15 November 2026 (Table 1). We asked five more models a subset. Each answer came from a single message sent through a programming interface, with no system prompt, no web search or other tools, and default settings. This measures the models themselves, not the assistants built on them.

4.1 The judge and the script

Claude Fable 5.2 graded every answer (Anthropic, 2026). It saw the question as asked, the official answer, the required points and the answer. It was not told which model wrote the answer. Its instructions require every listed point, stated for New Zealand. We fixed those instructions before any answer was graded, and wrote and froze the script before the answers were collected.

4.2 The human check

Two volunteers each graded the same random sample of 120 answers. They answered the same two questions as the judge, without seeing the judge's grade or which model wrote the answer. Half of the sample came from Claude models and half from other companies' models. Half were answers the judge had passed and half were answers it had failed. Where the two volunteers disagreed about an answer, we set it aside. The pass mark was set in advance: people had to agree with the judge on at least 85% of answers, weighted to represent all graded answers.

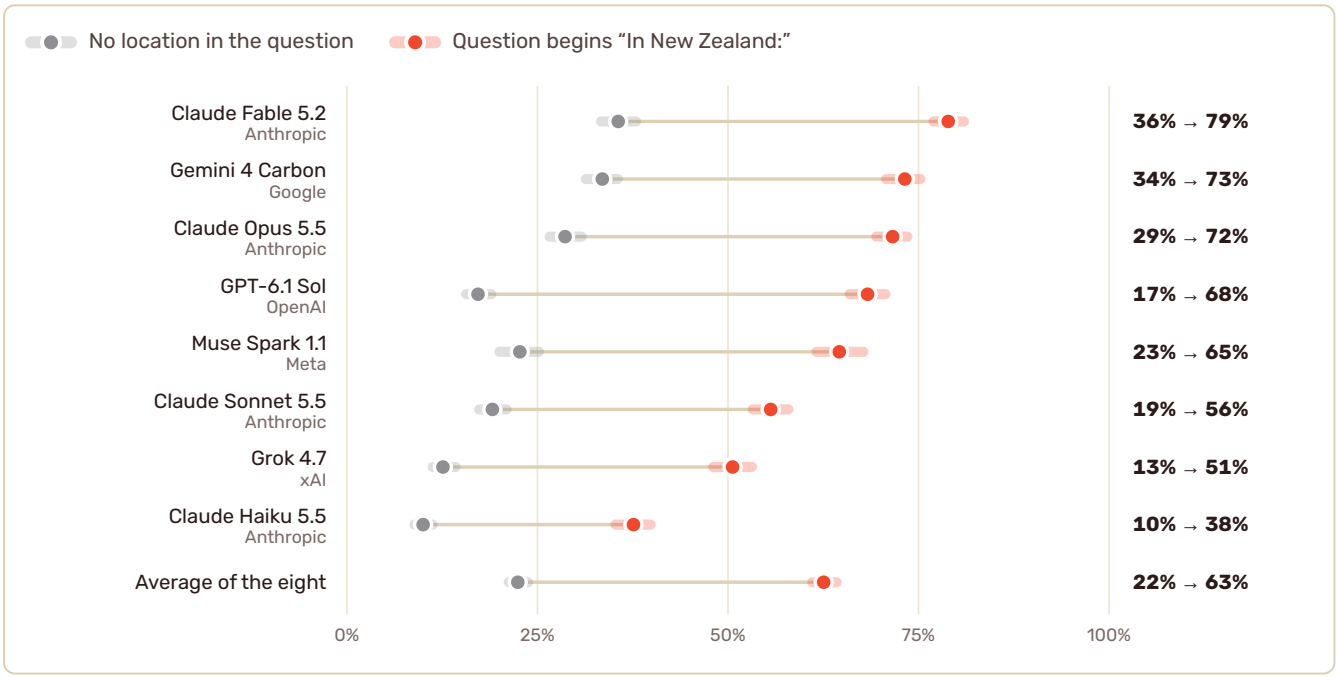
Statistics. The ranges we report show how much a result could move if we had asked a different set of questions. We calculate each range by resampling the questions many times and taking the middle 95% of the results, a method called the bootstrap (Efron & Tibshirani, 1993). We made some changes to the plan we registered before the study. The technical supplement lists every change from the registered plan.

5 Results

5.1 Saying where you are

When the question named New Zealand, the eight models gave a complete answer to 62.6% of questions on average. When it did not, they gave a complete answer to 22.4% (Figure 1, Table 1). Naming New Zealand raised every model's score, by between 27.6 and 51.1 points. When New Zealand was named, the weakest model, Claude Haiku 5.5, gave a complete answer to 37.6% of questions. The strongest, Claude Fable 5.2, gave a complete answer to 78.9%. The script, which only checks for the key figure, shows the same pattern with higher scores: the correct New Zealand figure appeared in 74.1% of answers when New Zealand was named, and in 29.9% when it was not. We also tested five more models on a smaller set of questions. When New Zealand was named, they gave a complete answer to between 31% and 50% of those questions.

Figure 1. Naming New Zealand lifted every model, by 28 to 51 points



Share of questions answered completely, by model, with and without "In New Zealand:" in the question. Bars show 95% ranges.

Table 1. Main results for the eight models that were asked every question

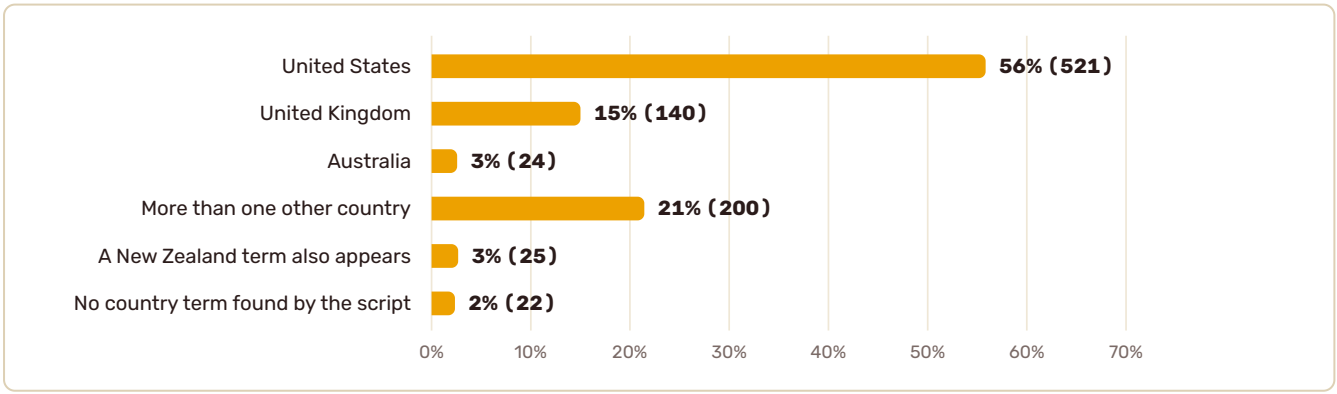
MODEL	QUESTIONS	COMPLETE, WITH LOCATION	COMPLETE, NO LOCATION	LOCATION EFFECT (POINTS)	ANOTHER COUNTRY, NO LOCATION	KEY FIGURE, WITH LOCATION	KEY FIGURE, NO LOCATION
Claude Fable 5.2	1,000	78.9% 76.3 to 81.6	35.6% 32.7 to 38.6	+43.3 +39.9 to +46.6	9.3% 7.6 to 11.2	86.4% 83.3 to 89.4	46.7% 42.3 to 51.2
Gemini 4 Carbon	1,000	73.2% 70.1 to 75.9	33.5% 30.7 to 36.2	+39.7 +36.4 to +42.8	10.1% 8.3 to 11.9	83.3% 80.3 to 86.4	43.7% 39.6 to 47.8
Claude Opus 5.5	999	71.6% 68.8 to 74.2	28.6% 25.9 to 31.4	+43.0 +40.0 to +46.3	5.6% 4.2 to 7.1	81.1% 77.4 to 84.3	38.4% 34.4 to 43.1
GPT-6.1 Sol	1,000	68.3% 65.3 to 71.3	17.2% 15.0 to 19.6	+51.1 +48.1 to +54.3	3.3% 2.2 to 4.6	81.5% 78.0 to 84.6	18.9% 15.6 to 22.4
Muse Spark 1.1	629	64.6% 61.0 to 68.4	22.7% 19.4 to 25.9	+41.8 +37.7 to +45.8	26.5% 23.3 to 30.0	77.3% 73.0 to 81.7	30.6% 25.7 to 35.6
Claude Sonnet 5.5	1,000	55.6% 52.6 to 58.6	19.1% 16.7 to 21.6	+36.5 +33.2 to +39.7	10.7% 8.9 to 12.7	68.1% 64.0 to 72.0	26.6% 22.6 to 30.3
Grok 4.7	1,000	50.6% 47.4 to 53.8	12.6% 10.6 to 14.9	+38.0 +34.7 to +41.4	20.9% 18.5 to 23.4	65.2% 61.2 to 69.3	18.3% 15.0 to 21.5
Claude Haiku 5.5	1,000	37.6% 34.6 to 40.5	10.0% 8.2 to 11.9	+27.6 +24.7 to +30.4	16.7% 14.4 to 19.0	50.0% 45.9 to 54.1	15.7% 12.6 to 19.1
Average of the eight		**62.6%** **60.4 to 64.9**	**22.4%** **20.6 to 24.4**	**+40.1**	**12.9%** **11.5 to 14.3**	**74.1%** **71.4 to 76.8**	**29.9%** **26.9 to 33.1**

Complete: the judge found every must include point, on all questions. Key figure: the script found New Zealand's figure, on the 508 figure questions. Figures in brackets are 95% ranges.

5.2 Which country an answer is written for

When the question did not have a location, 53.0% of answers hedged across several countries or asked where the person was. A further 12.9% were written for another country, ranging from 3.3% for GPT-6.1 Sol to 26.5% for Muse Spark 1.1. Where the script could name a single other country, it was the United States in 521 of 686 answers, the United Kingdom in 140 and Australia in 24 (Figure 2). Hedging is a reasonable response to a question that names no country, but it leaves the person without an answer. An answer for the wrong country is worse, because nothing in it says so.

Figure 2. When a model picked another country, it was usually the United States

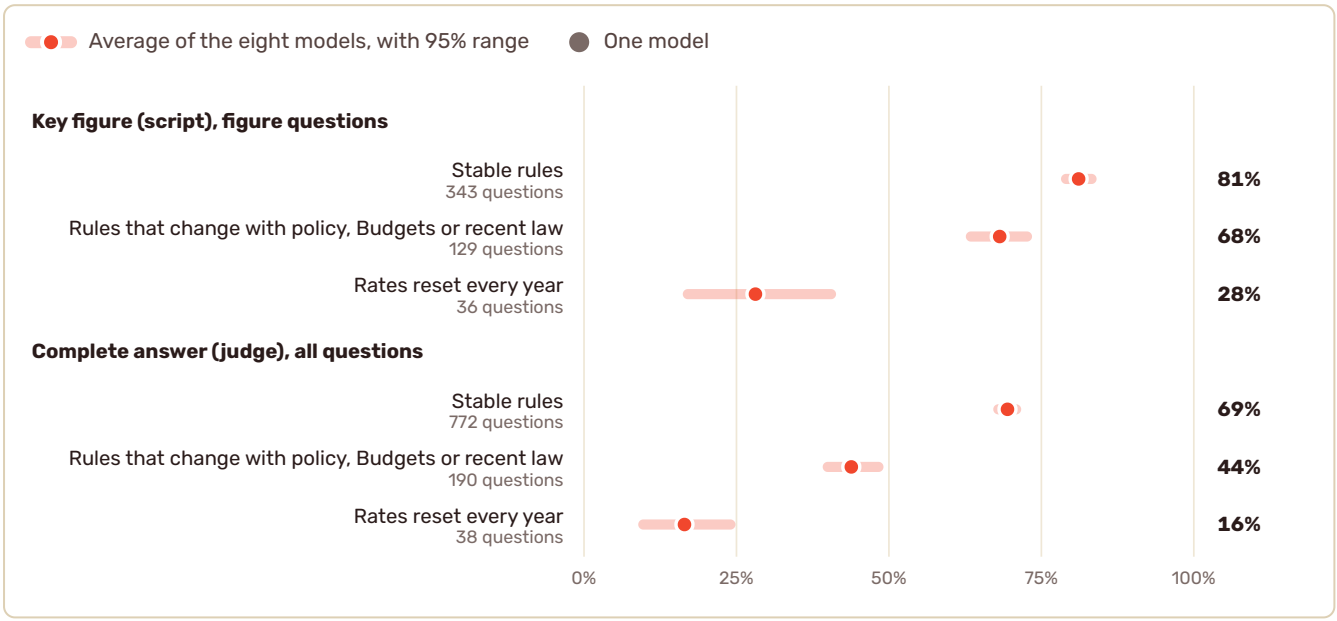


Answers to questions asked without a location that the judge found were written for another country, by the country the script identified. Not shown: Canada (1 answer).

5.3 Figures that change

Models were weakest where the rules change most often (Figure 3). With the location, the key figure appeared in 81.1% of answers on rules that rarely change. It appeared in 68.2% on rules that had recently changed or that change with Budgets, and in 28.1% on rates that are reset every year. On the 45 questions whose official figure had recently been replaced, 22.9% of answers gave only the old figure. A model can be right about the country and wrong about the year.

Figure 3. Answers were weakest where the rules change most often

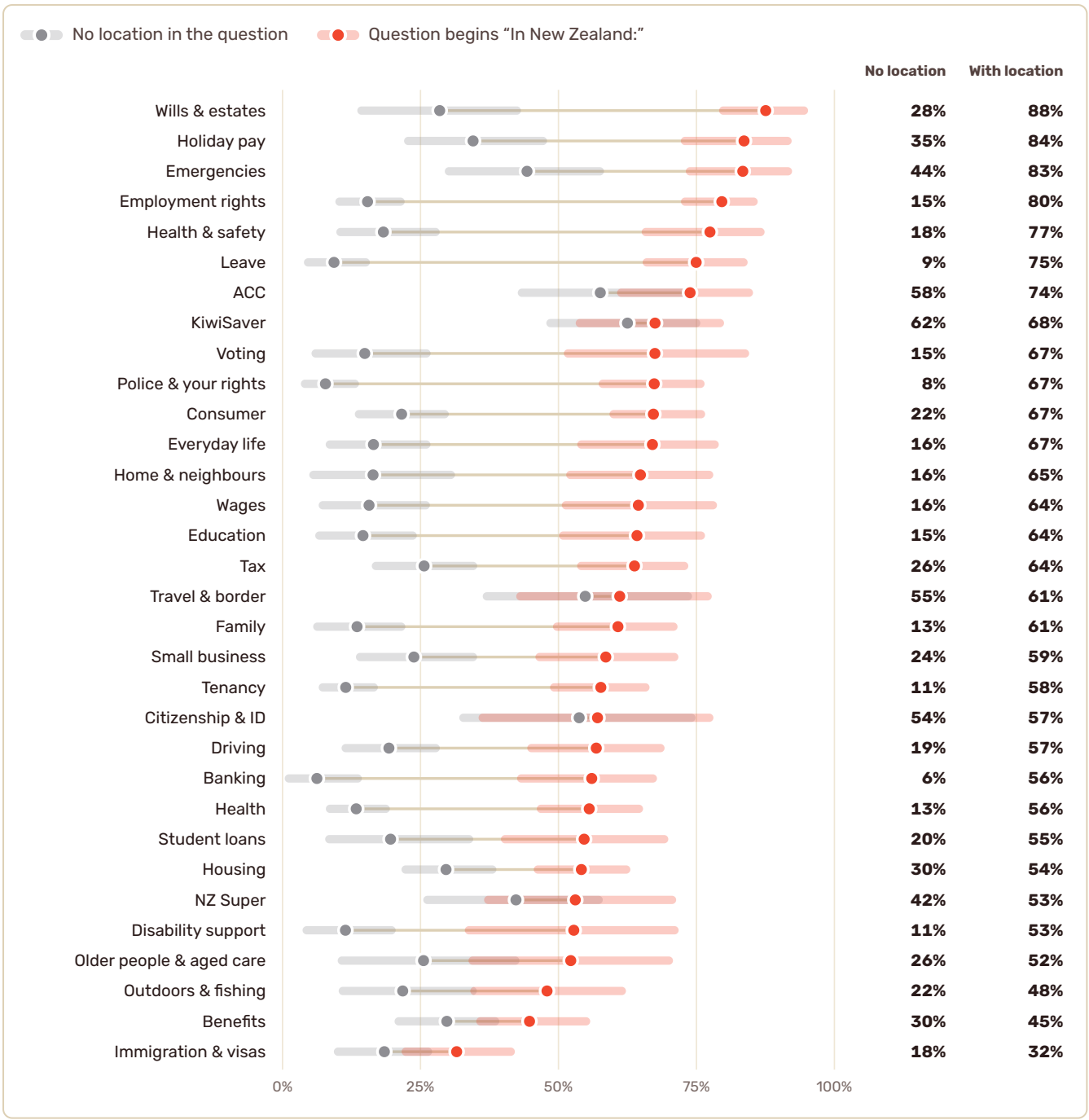


Share of answers giving New Zealand's key figure (script), with the location stated, by how often the rule changes.

5.4 Topics

Results varied widely by topic (Figure 4). The weakest topics were immigration and visas (32% complete), benefits (45%) and outdoors and fishing (48%). The strongest was wills and estates (88%). The same topics were hard for every model. This points to the subject matter, and to how often it changes, more than to any one model.

Figure 4. Complete answers ranged from 32% to 88% across topics

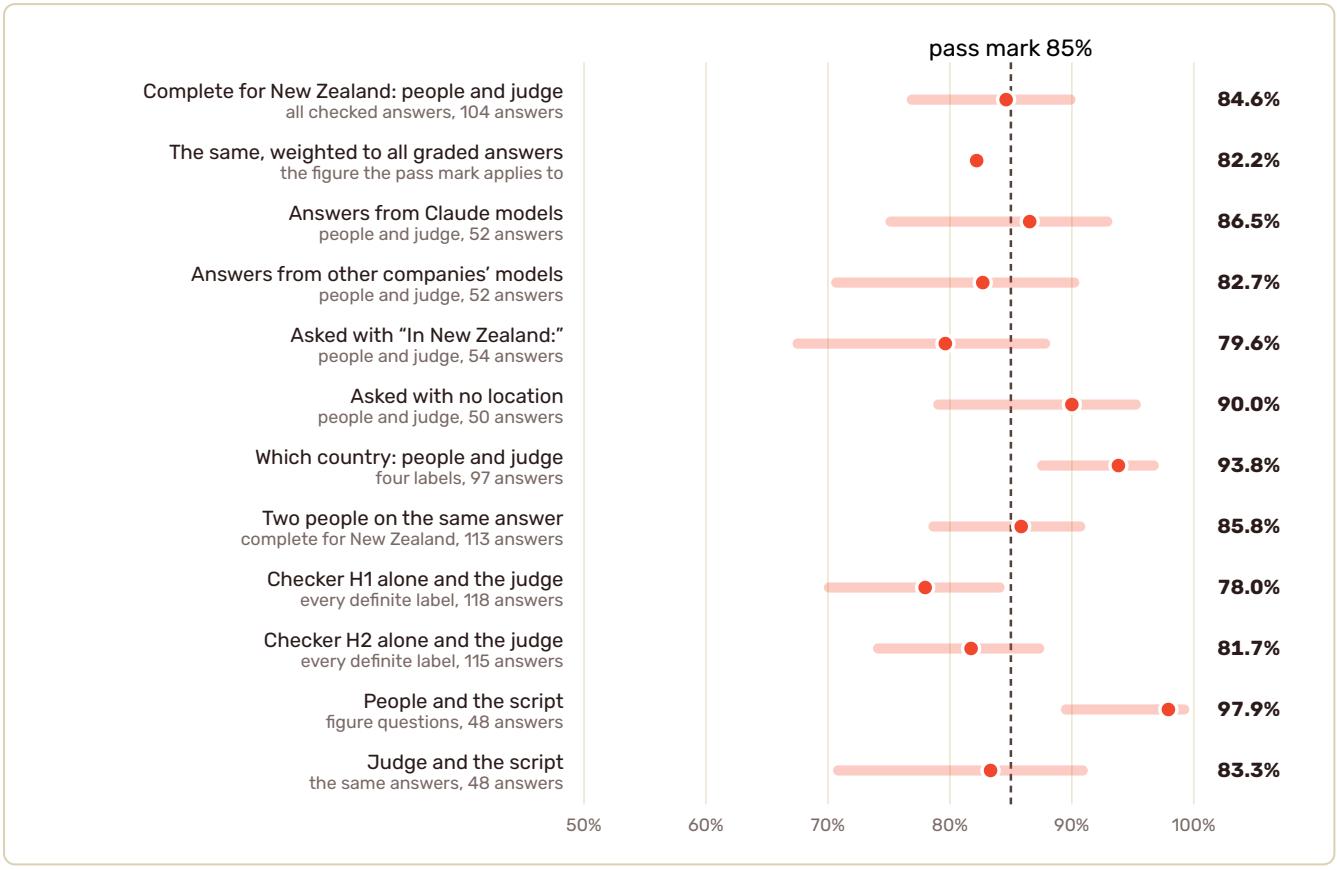


Share of questions answered completely with the location stated, by topic, averaged over the eight models. Bars show 95% ranges; topics with few questions have wide ranges.

6 Checking the judge

The judge did not meet its pass mark. The two checkers gave a clear label to 104 of the 120 sampled answers. They agreed with the judge on 84.6% of those, or 82.2% when weighted to represent all graded answers (kappa 0.68; Figure 5, Table 2). The pass mark was 85%. The specification allowed the judge's instructions to be revised once. The author chose not to do this, so the grades reported here are those of the registered judge. The two checkers agreed with each other on 85.8% of the answers they both labelled. Taken alone, one checker agreed with the judge on 78.0% of answers and the other on 81.7%.

Figure 5. People agreed with the judge on 84.6% of answers (82.2% weighted), under the 85% pass mark



Agreement on whether an answer was complete for New Zealand. The top row is the registered result: both checkers against the judge.

Table 2. The judge against people, and against the script

COMPARISON	ANSWERS COUNTED	AGREE	SHARE
Complete for New Zealand: people and judge	all checked answers	88 of 104	84.6%
Answers from Claude models	people and judge	45 of 52	86.5%
Answers from other companies' models	people and judge	43 of 52	82.7%
Asked with "In New Zealand:"	people and judge	43 of 54	79.6%
Asked with no location	people and judge	45 of 50	90.0%
Which country: people and judge	four labels	91 of 97	93.8%
Two people on the same answer	complete for New Zealand	97 of 113	85.8%
Checker H1 alone and the judge	every definite label	92 of 118	78.0%
Checker H2 alone and the judge	every definite label	94 of 115	81.7%
People and the script	figure questions	47 of 48	97.9%
Judge and the script	the same answers	40 of 48	83.3%

Registered figure: the first row, which weighted to all graded answers is 82.2%. Pass mark 85%: not met.

6.1 Why they disagreed

When the judge and people disagreed, the judge was always the stricter grader. People passed 16 of the 48 answers the judge had failed. They failed none of the 56 it had passed.

The script backs this up. On the sampled key figure questions people agreed with the script 97.9% of the time and the judge agreed with it 83.3% of the time.

The difference is in how strictly each applied the rule. The judge requires every listed point. People were given the same rule, but accepted answers that gave the main point.

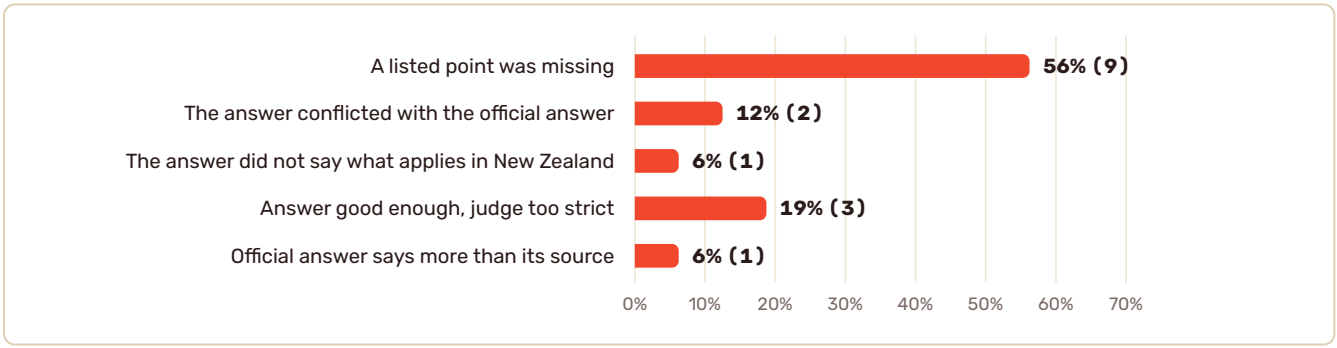
Once the check was finished the author read each of the 16 answers against its list of required points (Figure 6):

- In 12 the judge's reason held. A listed point was missing in 9. The answer conflicted with the official answer in 2. In 1 the answer did not say what applies in New Zealand.
- In 3 the answer was good enough and the judge was too strict.
- In 1 the official answer says more than its source does.

The technical supplement lists what was found in each answer.

Because the judge is strict, our complete answer scores are a minimum. The true figures are likely higher. The findings above do not rest on the judge alone. The location effect, the answers written for other countries, the weakness on changing figures and the order of the models all appear on the script as well.

Figure 6. Checked by hand, the judge's reason held in 12 of the 16 answers it failed and people passed



The author's reading of the 16 answers the judge failed and both checkers passed.

6.2 Does the judge favour its own model?

The judge is also the model with the most complete answers, and AI judges are known to favour their own output (Panickssery et al., 2024). We checked this in three ways. First, the script, which uses no AI, also puts that model first. Second, independent indices of general capability ranked it first in November 2026 (Artificial Analysis, 2026; Epoch AI, 2026). Third, people agreed with the judge on 86.5% of answers from Claude models and 82.7% of answers from other companies' models, which gives no sign that it was kinder to its own. We are more certain that this model leads than of how far ahead it is.

7 Limitations

The models, not the products. Answers came through programming interfaces, with no web search and no system prompt. Assistants that search the web, or that are told the user's location, may do better.

The judge missed its pass mark. The AI judge marked answers harder than people did. This means our scores for complete answers are probably lower than they should be. The script is our backup check but it only works on questions with a number or date as the answer.

A small human check. Two volunteers graded 120 answers. Their labels are a check on the judge. They are not the final word on any answer.

The judge and the leading model are the same. The questions were also drafted with models from the same family (Section 6.2). Questions drafted with one model family may suit that family in ways our checks cannot detect.

Answer keys. No lawyer has reviewed the official answers. They are quoted from official pages and were checked against those pages. The author's check of the contested answers found one official answer that says more than its source does. The benchmark holds no Māori data and makes no claim about te reo Māori or tikanga.

One moment in time. Rules change and models are replaced. These results are the November 2026 edition.

8 Conclusion

We asked widely used AI models everyday questions about New Zealand. When the question named New Zealand, they gave a complete answer approximately 3 out of 5 times. When New Zealand was not mentioned in the question this dropped to about one in every five answers. They were weakest on figures that are frequently updated or change. When they guessed a country, it was usually the United States. There are three simple ways to improve this. Add the location to the question. Give the model the current figure. Test an AI model before the public uses it. The judge that graded these answers missed its pass mark. It was stricter than people and never more lenient. Our main findings do not depend on it because the script shows the same results with no AI involved.

Ravenscar AI offers NZBench as a public measure of that gap: open to check, dated, and repeatable as the models and the rules change. In the author's view, discussion of AI safety in New Zealand has so far run ahead of the local means of testing AI systems. A benchmark of this kind is one practical part of those means.

Statements

Use of AI tools

The author conceived and directed the study, chose its topics and scope, and made each decision recorded in the registration files and the list of changes. The author also recruited the checkers, approved the spending and is responsible for the content of this report. AI tools were used throughout, under the author's direction. Claude models (Anthropic, 2026) were used to draft the questions and official answers from official pages under the rules in Section 3. They were also used to write the software, run the analyses, and prepare the figures and a first version of the text, which the author reviewed and revised. Claude Fable 5.2 was the judge (Section 4.1). It also made the first reading of the contested answers reported in Section 6.1, which the author then checked for the 16 answers the paper reports. The human check was done by people, who were asked not to use AI tools. Every figure from the study is read from a results file when the report is built.

No AI system is listed as an author. Authorship carries accountability that an AI system cannot take on (COPE Council, 2023), and the model is cited as a tool in the form APA style sets out (McAdoo et al., 2025).

Funding

The study was paid for by Ravenscar AI. Fees for the Claude models, including the judge, were met from credits that Ravenscar AI holds through Anthropic's programme for start-ups; fees for other companies' models were paid by Ravenscar AI. In total the work cost US$1,100.93 in API fees (about NZ$1,960). Anthropic API credits came to US$1,001.30 (about NZ$1,782). This covers an earlier pilot on 7 November 2026 and the study itself on 14 and 15 November including the judge. OpenRouter credits for the other companies' models came to US$99.63 (about NZ$177). The US dollar figures are the billed totals from the two providers' dashboards on 18 November 2026. The New Zealand dollar figures are estimates at NZ$1.78 to US$1 which was the mid-market rate in early November 2026. Anthropic had no part in the design, the analysis or the decision to publish.

Interests

The author is the founder of Ravenscar AI, which may offer paid services related to NZBench (the technical supplement). The benchmark was built with, and is judged by, models from Anthropic, one of the companies whose models are tested and the one whose model ranks first.

Data and code

Released with this report: the 924 public questions, official answers and sources; the models' answers; the judge's verdicts; the script's scores; the checkers' labels under codes (H1, H2 and so on); and all code, including the code that builds this report. The data are released under the CC BY 4.0 licence and the code under the MIT licence, at https://doi.org/10.5281/zenodo.24364019 (DOI 10.5281/zenodo.24364019), and linked from https://ravenscar.ai/nzbench.

Not released: the 76 held back questions, with their answers and the models' replies to them; the checkers' names; and, until the terms of reuse are confirmed, the word for word passages from the Community Law Manual, for which the address of the source is given instead.

Ethics

The study collected no personal information. Its questions describe invented situations. Any resemblance to a personal situation is merely coincidence. The two checkers were adult volunteers who agreed to take part and are not named.

The question bank holds no Māori data and makes no claim about te reo Māori or tikanga. An edition that covered them would need to be led with Māori partners and governed in keeping with the principles of Māori data sovereignty (Te Mana Raraunga, 2018).

Acknowledgements

Thanks to the checkers, and to the agencies and Community Law, whose public information the official answers quote. They have not reviewed this report and are not responsible for anything in it.

Citing and contact

Suggested citation: Carter-Scofield, B. (2026). *Does AI get New Zealand right? NZBench, November 2026 edition* [Research report]. Ravenscar AI. https://doi.org/10.5281/zenodo.24364019

Corrections to an official answer, and requests to use the benchmark, are welcome at hello@ravenscar.ai.

The technical supplement is the full NZBench report (available at ravenscar.ai/nzbench): the complete method, every result and table, the changes from the registered plan, what was found in each contested answer, and the datasheet for the question bank.

References

Anthropic. (2026). *Claude Fable 5.2* [Large language model]. https://www.anthropic.com

Artificial Analysis. (2026). *LLM leaderboard: Comparison of AI models*. Retrieved November 17, 2026, from https://artificialanalysis.ai/leaderboards/models

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short papers)* (pp. 238–255). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-short.20

COPE Council. (2023). *COPE position: Authorship and AI*. Committee on Publication Ethics. https://doi.org/10.24318/cCVRZBms

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93. https://doi.org/10.1093/jla/laae003

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Epoch AI. (2026). *Claude Fable 5.2*. Retrieved November 17, 2026, from https://epoch.ai/models/claude-fable-5-2

Etxaniz, J., Azkune, G., Soroa, A., Lopez de Lacalle, O., & Artetxe, M. (2024). *BertaQA: How much do language models know about local culture?* arXiv. https://arxiv.org/abs/2406.07302

InternetNZ. (n.d.). *New Zealand's internet insights: 2025 survey findings*. Retrieved November 16, 2026, from https://internetnz.nz/assets/Archives/New-Zealands-Internet-Insights-2025-survey-findings.pdf

Kasai, J., Sakaguchi, K., Takahashi, Y., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., & Inui, K. (2022). *RealTime QA: What's the answer right now?* arXiv. https://arxiv.org/abs/2207.13332

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., & Blunsom, P. (2021). Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 29348–29363). Curran Associates.

Lee, J., Kim, M., Kim, S., Kim, J., Won, S., Lee, H., & Choi, E. (2024). *KorNAT: LLM alignment benchmark for Korean social values and common knowledge*. arXiv. https://arxiv.org/abs/2402.13605

Libovický, J., Helcl, J., Manea, A., & Vico, G. (2025). *CUS-QA: Local-knowledge-oriented open-ended question answering dataset*. arXiv. https://arxiv.org/abs/2507.22752

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242. https://doi.org/10.1111/jels.12413

Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). *Large language models are geographically biased*. arXiv. https://arxiv.org/abs/2402.02680

McAdoo, T., Denneny, S., & Lee, C. (2025, September 9). *Citing generative AI in APA Style: Part 1—Reference formats*. APA Style. https://apastyle.apa.org/blog/cite-generative-ai-references

Miller, E. (2024). *Adding error bars to evals: A statistical approach to language model evaluations*. arXiv. https://arxiv.org/abs/2411.00640

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez-García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., . . . Oh, A. (2024). *BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages*. arXiv. https://arxiv.org/abs/2406.09948

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Panickssery, A., Bowman, S. R., & Feng, S. (2024). *LLM evaluators recognize and favor their own generations*. arXiv. https://arxiv.org/abs/2404.13076

Prior, M., Schultz, A., & Grabmair, M. (2026). *Asking for an old friend: Diagnosing and mitigating temporal failure modes in LLM-based statutory question answering*. arXiv. https://arxiv.org/abs/2605.23497

Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A. G., & Arawjo, I. (2024). Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *UIST '24: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery. https://doi.org/10.1145/3654777.3676450

Te Mana Raraunga. (2018). *Principles of Māori data sovereignty* (Brief No. 1). https://www.temanararaunga.maori.nz/

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., & Luong, T. (2024). FreshLLMs: Refreshing large language models with search engine augmentation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 13697–13720). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.813

Wang, Z., & Suresh, H. (2026). *Which institutional frameworks do chatbots assume? Auditing jurisdictional defaults in multilingual LLMs*. arXiv. https://arxiv.org/abs/2606.00333

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. arXiv. https://arxiv.org/abs/2306.05685