

# Voice of Australian-AI: Learning Australian Assistant Voice Without Caricature

Yue Yang<sup>1</sup>, Maxwell Twelftree<sup>1</sup>, Luke Borgnolo<sup>1</sup>,  
David Lemphers<sup>1</sup>, Amanda Baughan<sup>1</sup>,

<sup>1</sup>Maincode, Melbourne, Australia,

Correspondence: [yue/amanda@maincode.com](mailto:yue/amanda@maincode.com)

## Abstract

Interactional defaults in widely deployed LLM assistants tend to align with US conventions, including in non-US contexts. We introduce an Australian-facing assistant, Matilda, and test whether local voice is better learned during fine-tuning or imposed with an inference-time prompt. We define Australian Voice as pragmatic alignment, directness, modesty, warmth, register control, and task fidelity, rather than slang or spelling substitution. We build a governed dataset from user research, human rewrites, contrastive negatives, reviewed synthetic examples, and replay data. In a blind evaluation by Australian English users (356 question-level ratings over 89 prompts), the raw 50% voice/replay mix scores best on all five dimensions (4.24/5; 102 best-pick wins). Prompted variants underperform every matched fine-tuned model.

## 1 Introduction

LLM assistants are increasingly deployed across English-speaking markets, yet English fluency does not imply cultural or pragmatic fit. A response can be factually acceptable and still feel over-formal, promotional, evasive, unserious, or culturally mismatched (Karinshak et al., 2024; Qiu et al., 2025). For an industry assistant, this is not a cosmetic problem. This affects whether users trust an answer, whether they follow instructions, and whether they regard the system as suitable for ordinary work, customer support, and life-administration tasks.

Australian English makes this problem concrete. Australian users often prefer assistant responses that are practical, direct, modest, warm, and low-ceremony; while rejecting exaggerated slang or broad national performance (Cox and Paethorpe, 2007). Simply inserting tokens such as “mate”, “reckon”, or “yeah nah” can ironically make a response feel less Australian when the usage is forced, poorly timed, or inappropriate for the setting. Conversely, a response can contain no overt

local markers yet still be perceived as Australian if it gets stance, register, and pragmatic expectations right.

We therefore frame *Australian Voice* as interactional cultural alignment. The target is not accent imitation, dense slang, or spelling transfer. It is a deployable response behaviour: know when to be casual, when to be precise, when to avoid humour, when to use plain wording, and when local context matters for task fidelity. This framing connects cultural alignment work in NLP (Durmus et al., 2023; Liu et al., 2023; Montalan et al., 2024; Ke et al., 2026; Yang et al., 2025) with sociolinguistic accounts of Australian English as a socially meaningful English dialect (Wierzbicka, 1986; Goddard, 2009; Goddard and Cramer, 2016; Collins, 2012; Coats, 2022, 2024; Diskin-Holdaway and Escudero, 2024).

This paper presents Australian-AI as an industry case study in learning such a voice. We construct Australian Voice data from user interviews, a written voice specification, human rewrites, negative contrastive examples, reviewed synthetic examples, and general replay data. We then fine-tune model variants with different mixture weights and compare them with prompt-engineered variants that attempt to impose Australian style at inference time.

The central empirical result is simple: in blind human evaluation, learned voice beats prompted voice. The raw 50% Australian Voice mixture is the strongest system overall, while prompt-engineered variants lose to their matched raw fine-tuned models at 0%(base model), 30%, and 50% mixture levels. This matters because prompt-level localisation is attractive in production: it is cheap, reversible, and easy to describe. Our results suggest that this convenience comes at a quality cost. When the desired behaviour is pragmatic and context-sensitive, the model should learn it from curated examples rather than perform it from a visible style instruction.

Our contributions are fourfold. First, we operationalise Australian Voice as a measurable alignment target with explicit anti-caricature, register-control, and task-fidelity requirements. Second, we describe a governed data pipeline that combines user-grounded human rewrites, contrastive negatives, reviewed synthetic distillation, and general replay data. Third, we provide a controlled human comparison of raw fine-tuned and prompted variants across three mixture weights. Fourth, we separate the main voice result from a broader capability screen, treating benchmark results as regression checks rather than as the paper’s primary claim.

## 2 Related Work

Cultural alignment work argues that LLM behaviour reflects the assumptions embedded in training data, evaluation data, and alignment procedures. Existing models may represent some countries’ opinions and social assumptions more strongly than others, and they can struggle with culturally situated reasoning even when the language is fluent (Durmus et al., 2023; Liu et al., 2023; Montalan et al., 2024). This motivates evaluation beyond factual correctness: a model may answer fluently while missing local pragmatic expectations.

Australian English is a national variety with lexical, phonetic, and social meaning. Its cultural force is not limited to a vocabulary list. Work on Australian English highlights themes such as informality, understatement, irreverence, and local social identity (Wierzbicka, 1986; Goddard, 2009; Goddard and Cramer, 2016; Collins, 2012; Coats, 2022, 2024; Diskin-Holdaway and Escudero, 2024). For assistant alignment, this suggests that the important unit is not “Australian words” but interactional stance: how direct, modest, humorous, careful, or practical a response should be in context.

## 3 Australian Voice as an Alignment Target

Australian English is a national variety with lexical, phonetic, and social meaning. Its cultural force is not limited to a vocabulary list. Work on Australian English highlights themes such as informality, understatement, irreverence, and local social identity (Wierzbicka, 1986; Goddard, 2009; Goddard and Cramer, 2016; Collins, 2012; Coats, 2022). We treat localisation as a constrained behaviour-learning problem. Given a user request  $x$ , the model should produce a response  $y$  that satisfies

the task while matching Australian interactional expectations. The target is context-dependent: casual planning may allow warmth or light informality, while technical, workplace, complaint, emotional, or safety-sensitive requests require steadier register control (Coats, 2024; Diskin-Holdaway and Escudero, 2024). Local voice is therefore not a global style, but a judgement about how to sound while still doing the job. Our voice specification was derived from interviews with Australian product users and review of Australian English reference material. It targets mainstream contemporary Australian English without stereotype: direct, practical, modest, warm, and low-ceremony. The desired stance is a capable peer rather than a lecturer, salesperson, or servant: get to the point early, use plain wording, favour understatement over hype, and provide warmth without excessive enthusiasm. This is a product-language target, not a linguistic claim about all Australian English speakers.

The specification also defines what the assistant should *not* do. It forbids phonetic accent spelling, dense slang, broad “ocker”, or crude, uncultivated speech, gratuitous banter, and replacing useful content with local colour. It also avoids adopting or simulating Aboriginal English or other community and regional varieties. These constraints make the evaluation target explicit: a response should improve Australian naturalness and cultural fit without sacrificing restraint, register control, helpfulness, factuality, or safety. A model that becomes more slang-heavy, less accurate, too casual in serious settings, or friendly while missing the practical next step has failed. In this sense, Australian Voice is closer to alignment than to style transfer.

## 4 Australian Voice Data Pipeline

We curate an Australian Voice dataset to better fit the communication requirements and expectations of Australian users. The dataset combines positive examples, negative examples, reviewed synthetic examples, and replay data. As shown in Figure 1, the pipeline starts from user-defined requirements for the target Australian voice, including tone, register, phrasing, directness, cultural fit, and interaction style. Australian annotators then rewrite and review examples to ensure that the responses sound natural to local users rather than simply using generic English. We further use synthetic generation to expand coverage across more domains and scenarios, while manual review keeps the gen-



Figure 1: Australian Voice data curation pipeline.

erated data aligned with the desired voice. Finally, replay data is mixed in to preserve the model’s general capabilities and reduce over-specialisation to Australian-style responses.

The pipeline has eight stages: (1) user interviews, (2) voice specification, (3) reference-material review, (4) human rewrite collection, (5) few-shot seed construction, (6) synthetic data distillation, (7) Australian reviewer filtering, and (8) iterative prompt/specification refinement.

#### 4.1 Dataset Construction and Curation

**Few-shot seeds: positive and negative supervision.** We first constructed a set of few-shot seed examples from topics that reflect the everyday use cases and communication requirements of the target Australian user base. For each topic, we generated 20 seed prompts covering realistic user requests, including customer communication, workplace writing, local explanations, technical assistance, life administration, and serious or high-stakes contexts. For each selected prompt, we sampled an initial response from the base model and asked Australian annotators to rewrite it according to the Australian Voice specification. The rewrite task was deliberately constrained. Annotators were instructed to preserve the original task intent, factual content, reasoning, correctness, and safety behaviour, while adapting only the response style: tone, register, phrasing, directness, warmth, and degree of localisation. As a result, approved rewrites are treated as preferred behaviour examples rather than surface-level paraphrases. They teach the model how an answer should sound when it is already substantively correct. To make the boundary of the voice explicit, we also created contrastive negative examples. These examples capture common failure modes of superficial localisation, including excessive slang, forced humour, US-

style customer-service exuberance, British-style stiffness, corporate hype, exaggerated “ocker” performance, irrelevant banter, inappropriate humour in serious contexts, and hallucinated local details. The positive and negative examples therefore provide both attraction and repulsion signals: the model learns the desired Australian voice, while also learning which forms of localisation should be avoided.

**Reviewed synthetic expansion.** After collecting the human-written seed examples, we used them as few-shot demonstrations for a stronger teacher model to generate additional Australian Voice training data. This expansion step increased coverage across domains, interaction types, and levels of required localisation. The synthetic prompts and responses covered customer support, complaints, workplace communication, local explanations, casual planning, serious contexts, and style-control cases where Australianisation should be light, neutral, or absent. Synthetic examples were not added to the training set automatically. Each generated item was reviewed by Australian annotators, who either approved it, edited it, or rejected it. Reviewers checked whether the response preserved the user’s intent, maintained factual accuracy and safety behaviour, used an appropriate register, and matched the Australian Voice specification without becoming performative. Common rejection reasons included overuse of slang, excessive informality, unnecessary banter, hallucinated local references, loss of factual content, inappropriate humour, and register mismatch. This review step allowed synthetic generation to improve coverage while keeping the final dataset aligned with the intended user experience.

#### 4.2 Data Format

We store the dataset in a standard supervised fine-tuning format. Each example contains a question, an answer, and a category. The question field contains the user prompt, the answer field contains the preferred model response, and the category field records the topic or use-case type of the example. This simple format makes the dataset easy to mix with replay data, filter by domain, and use directly for supervised instruction tuning.

question : user prompt,  
 answer : preferred Australian Voice response,  
 category : topic or use-case category

The current Australian Voice dataset contains ap-

Table 1: Core data signals used for Australian voice alignment.

| Signal                        | Role                                                                                                                             |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| User research and voice guide | Defines the target stance: direct, modest, warm, practical, restrained, and context-sensitive.                                   |
| Human rewrites                | High-precision preferred responses reviewed for voice fit, task fidelity, non-caricature, and register control.                  |
| Contrastive negatives         | Examples of what not to learn: forced slang, caricature, hype, irrelevant banter, and local hallucination.                       |
| Reviewed synthetic data       | Coverage expansion across support, complaints, workplace tasks, local explanations, serious contexts, and style-control prompts. |
| General data                  | Preserves broad assistant capability and prevents every response from becoming a local identity performance.                     |

proximately 20,000 extracted and curated question–answer pairs. These examples cover everyday Australian assistant use cases, including workplace writing, customer and business communication, civic and financial explanation, technical support, personal administration, and plain-language rewriting. Each accepted pair is stored in a supervised fine-tuning format with a user question, a preferred Australian Voice answer, and a topic category.<sup>1</sup>

## 5 Australian Voice Model Training

We fine-tune the model using supervised instruction tuning on a mixture of Australian Voice data and general replay data. Let  $\mathcal{D}_{\text{AU}}$  denote the Australian Voice dataset and  $\mathcal{D}_{\text{G}}$  denote the general replay dataset. For each experiment, we construct a mixed training distribution

$$\mathcal{D}_{\text{mix}} = \lambda \mathcal{D}_{\text{AU}} + (1 - \lambda) \mathcal{D}_{\text{G}},$$

where  $\lambda$  controls the proportion of Australian Voice data used during fine-tuning. To study the effect of Australian Voice exposure, we train five mixture settings:  $\lambda \in \{0.10, 0.20, 0.30, 0.50, 1.00\}$ . These correspond to 10%, 20%, 30%, 50%, and 100% Australian Voice mixtures. Lower values of  $\lambda$  preserve more general replay data, while higher values expose the model more strongly to Australian Voice examples. The 100% setting removes replay data entirely and serves as a stress-test for whether full specialization improves local voice quality or harms broader model capability.

<sup>1</sup>The dataset is currently used for internal model training and evaluation. We are also working with participating contributors and the company to finalise the consent, licensing, and release agreements required to make a cleaned version of the dataset publicly available.

For each training instance with input context  $x$  and target assistant response  $y$ , we minimise assistant-token negative log likelihood:

$$\mathcal{L}(\theta) = - \sum_{(x,y) \in \mathcal{D}_{\text{mix}}} \sum_{t=1}^{|y|} \log p_{\theta}(y_t \mid x, y_{<t}).$$

The mixture design reflects a production trade-off. Increasing  $\lambda$  is expected to improve Australian naturalness, idiom use, and local register alignment. However, overly large Australian Voice proportions may over-specialise the model, reduce general capability, or introduce unnecessary stylistic shifts in tasks where a neutral response is preferable. General replay data therefore acts as a retention mechanism: it helps preserve broad instruction-following, reasoning, coding, factuality, and tool-use behaviour while still allowing the model to acquire Australian Voice characteristics.

## 6 Evaluating Capability Retention and Australian Voice Quality

We evaluate the Australian Voice model along two dimensions. First, we measure whether fine-tuning preserves the model’s general capabilities on standard benchmark tasks, including reasoning, mathematics, knowledge, instruction following, coding, and tool-use benchmarks. This checks that adapting the model to Australian user requirements does not substantially degrade its broader task performance. Second, we conduct a human evaluation focused on Australian Voice quality. Australian reviewers compare model responses and assess whether they sound natural, appropriate, and useful for Australian users. This evaluation measures aspects that are difficult to capture with automatic benchmarks, such as tone, register, directness, cultural fit, restraint, and avoidance of forced or exaggerated localisation. Together, these evaluations allow us to identify the best mixture weight for deployment. An ideal model should improve Australian Voice alignment while maintaining strong performance on general-purpose benchmarks. retention.

### 6.1 General Capability Evaluation

For general capability, we run a broad benchmark suite covering knowledge, reasoning, instruction following, dialogue quality, coding, and tool-use. The benchmark set includes MMLU-Pro, MATH-500, AIME, SimpleQA, IFEval, Arena-Hard, MT-Bench, WildBench, LiveBench, BigCodeBench,

Table 2: Fine-tuned evaluation matrix across 4B, 8B, 14B, and 35B models. Each numeric cell is a pass rate in percent, and bold marks the best available result within each size class and benchmark. Variants: A is 10%, B is 20% mix, C is 30%, D is 50%, E is 100% voice-strong

|     |      | Tier 1 (deterministic) |             |             |             |             | Tier 2 (LLM judge) |              |             |             | Tier 3 (custom) |             |             |             |               |
|-----|------|------------------------|-------------|-------------|-------------|-------------|--------------------|--------------|-------------|-------------|-----------------|-------------|-------------|-------------|---------------|
|     |      | MMLU-Pro               | MATH-500    | AIME        | SimpleQA    | IFEval      | Arena-H            | MT-Bench     | WildB.      | LiveB.      | BigCode         | HE+         | MBPP+       | BFCL        | $\tau$ -bench |
| 35B | A    | 97.5                   | 93.4        | <b>63.3</b> | <b>13.3</b> | 91.9        | 96.2               | 96.2         | 79.7        | 67.7        | 58.8            | <b>93.3</b> | 89.6        | 91.2        | <b>5.5</b>    |
|     | B    | 97.5                   | 93.0        | <b>63.3</b> | 12.8        | <b>91.3</b> | 96.2               | 98.8         | 85.9        | <b>78.1</b> | 56.7            | 90.9        | 69.8        | <b>93.8</b> | 4.8           |
|     | C    | <b>97.6</b>            | <b>94.0</b> | <b>63.3</b> | 12.4        | 96.2        | <b>97.6</b>        | <b>100.0</b> | <b>86.7</b> | 67.7        | <b>60.1</b>     | 92.7        | <b>70.1</b> | 88.3        | <b>5.5</b>    |
|     | D    | <b>97.6</b>            | 91.4        | 60.0        | 12.4        | 96.2        | 96.1               | <b>100.0</b> | <b>86.7</b> | 67.7        | <b>60.1</b>     | 92.7        | <b>70.1</b> | 88.3        | <b>5.5</b>    |
|     | E    | 80.6                   | 77.6        | 37.3        | 3.4         | 73.3        | 77.8               | 67.1         | 69.7        | 56.4        | 56.7            | 92.7        | 63.5        | 68.4        | 4.8           |
|     | base | <b>97.6</b>            | 91.9        | 66.7        | 14.3        | 91.9        | 93.1               | <b>98.8</b>  | <b>85.9</b> | 67.7        | <b>60.1</b>     | 92.7        | <b>70.1</b> | 88.3        | <b>5.5</b>    |
| 14B | A    | 80.1                   | <b>92.0</b> | 60.0        | 3.2         | 54.9        | 77.8               | 82.5         | 50.9        | 13.6        | 4.4             | 78.0        | 63.5        | 68.4        | 4.8           |
|     | B    | 82.9                   | 91.4        | 60.0        | 3.2         | 59.0        | 77.8               | 82.5         | 50.5        | 13.9        | <b>13.1</b>     | 78.0        | 64.8        | 69.1        | <b>5.5</b>    |
|     | C    | 80.6                   | 91.2        | 60.0        | 3.4         | 56.6        | 78.4               | <b>88.8</b>  | <b>52.7</b> | 14.5        | 10.2            | 79.3        | <b>69.0</b> | 69.1        | <b>5.5</b>    |
|     | D    | 84.7                   | 91.8        | 60.0        | 3.2         | 61.7        | <b>80.4</b>        | 86.2         | 50.4        | 15.4        | 3.4             | 77.4        | 60.1        | 67.7        | <b>5.5</b>    |
|     | E    | 82.8                   | 77.8        | 26.7        | 2.9         | 63.6        | 64.0               | 75.0         | 13.0        | 8.5         | 2.1             | 78.7        | 62.7        | 71.2        | 4.8           |
|     | base | <b>97.1</b>            | 85.2        | <b>63.3</b> | 2.2         | <b>73.0</b> | 16.6               | 7.5          | 35.6        | <b>21.4</b> | 0.1             | <b>86.0</b> | 64.6        | <b>79.3</b> | <b>5.5</b>    |
| 8B  | A    | 93.8                   | 77.6        | 20.0        | 2.5         | 64.1        | 41.8               | 18.8         | 11.8        | 8.4         | 41.3            | 69.5        | 56.6        | 67.0        | 4.2           |
|     | B    | 94.3                   | 76.6        | <b>33.3</b> | 2.5         | 61.0        | 44.2               | 25.0         | 11.9        | 8.4         | <b>49.7</b>     | 66.5        | 56.3        | 67.4        | 4.2           |
|     | C    | 94.5                   | 79.6        | 20.0        | <b>2.9</b>  | 59.7        | 36.0               | 18.8         | 8.8         | 7.0         | 46.8            | 69.5        | 56.3        | 66.1        | 4.8           |
|     | D    | 86.2                   | 70.6        | 16.7        | 2.8         | 62.8        | <b>58.6</b>        | <b>36.2</b>  | 13.6        | 7.0         | 0.7             | 74.4        | <b>61.4</b> | 71.0        | <b>5.5</b>    |
|     | E    | 84.2                   | 67.6        | 16.7        | 2.5         | 61.0        | 23.4               | 8.8          | 12.6        | 7.2         | 11.4            | 72.0        | 59.3        | 66.6        | 4.8           |
|     | base | <b>97.1</b>            | <b>81.8</b> | 20.0        | 1.8         | <b>71.3</b> | 16.0               | 15.0         | <b>41.9</b> | <b>21.3</b> | 0.2             | <b>80.5</b> | 60.1        | <b>82.1</b> | <b>5.5</b>    |
| 4B  | A    | 75.2                   | 77.8        | 20.0        | 2.2         | 57.3        | 23.8               | 15.0         | 12.2        | 7.2         | 58.9            | 54.9        | 51.1        | 58.6        | <b>5.5</b>    |
|     | B    | 75.2                   | 77.8        | <b>26.7</b> | 2.1         | 57.5        | 17.8               | 12.5         | 12.5        | 8.9         | 57.2            | 58.5        | 56.3        | 63.4        | 4.8           |
|     | C    | 80.4                   | <b>78.6</b> | <b>26.7</b> | <b>2.4</b>  | 58.6        | <b>36.8</b>        | 20.0         | 8.9         | 6.4         | <b>63.8</b>     | 64.0        | 58.2        | 64.0        | 4.8           |
|     | D    | 80.3                   | 76.2        | 13.3        | 2.2         | 65.1        | 30.0               | 15.0         | 10.8        | 7.0         | 48.9            | 56.1        | <b>60.1</b> | 61.6        | <b>5.5</b>    |
|     | E    | 80.4                   | 72.6        | 13.3        | 1.7         | 55.4        | 19.4               | 11.2         | 10.1        | 11.2        | 46.1            | 67.7        | 54.8        | 58.8        | 4.8           |
|     | base | <b>95.2</b>            | 78.2        | 13.3        | 1.2         | <b>72.1</b> | 10.2               | 18.8         | <b>47.5</b> | <b>26.5</b> | 53.5            | <b>72.0</b> | 59.5        | <b>68.9</b> | <b>5.5</b>    |

HumanEval+, MBPP+, and BFCL. This evaluation measures whether Australian Voice fine-tuning preserves the base model’s non-local capabilities across standard tasks. Table 2 shows that Australian Voice tuning can be added without materially degrading general capability when the mixture is kept moderate. For the 35B model, the 10–50% mixtures remain close to the base model across deterministic, LLM-judge, and custom benchmarks, with the 30% and 50% variants matching or exceeding the base model on several instruction-following and judge-based evaluations. In contrast, the 100% voice-strong model shows a clear capability drop. This supports our choice to evaluate the base, 30%, and 50% models in the human Australian Voice study: they provide the strongest practical trade-off between voice adaptation and general capability

## 6.2 Human Preference Study

For Australian Voice quality, we conduct a human study. Human annotators compare outputs from different model variants and judge whether the responses sound natural, locally appropriate, and recognisably Australian, without becoming forced, exaggerated, or stereotyped. We evaluate three main model families: the base model, the 30% Australian Voice mixture model, and the 50% Australian Voice mixture model. We exclude the 100%

Australian Voice setting from the human study because its general benchmark performance drops too substantially, making it unsuitable as a practical deployment candidate. We also exclude the 10% and 20% mixture models because their general performance is close to the 30% setting, and the 30% model provides a stronger test of whether Australian Voice adaptation is visible to human users. To examine the role of prompt engineering, which is commonly used in industry deployments, we evaluate each model in two conditions: with and without a short system prompt describing the desired Australian Voice. This allows us to test whether prompt engineering alone is sufficient for producing the target voice, or whether fine-tuning provides additional benefits beyond prompting.

We constructed 89 prompts spanning six usage areas: Australian civic/legal/finance topics, customer and business communication, technical/product/code tasks, workplace and team-management scenarios, personal/wellbeing/life-admin queries, and plain-language/educational/creative tasks. Australian native speakers or long-term Australian English users evaluated anonymised responses in a blind, randomised interface. Raters did not see model identities, and response order was shuffled. For each prompt, raters scored each candidate

Table 3: Overall mean score by aspect and training mix.

| Training mix     | Naturalness | Fit         | Restraint   | Register    | Helpfulness | Overall     | Wins       |
|------------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|
| base raw         | 3.94        | 4.03        | 4.23        | 4.14        | 4.05        | 4.08        | 74         |
| base + prompt    | 3.37        | 3.57        | 4.16        | 4.03        | 3.87        | 3.80        | 28         |
| 30% mix raw      | 4.01        | 4.01        | 4.32        | 4.17        | 4.07        | 4.12        | 60         |
| 30% mix + prompt | 3.50        | 3.71        | 4.29        | 4.09        | 4.03        | 3.93        | 32         |
| 50% mix raw      | <b>4.22</b> | <b>4.21</b> | <b>4.42</b> | <b>4.25</b> | <b>4.13</b> | <b>4.24</b> | <b>102</b> |
| 50% mix + prompt | 3.54        | 3.73        | 4.35        | 4.15        | 3.78        | 3.91        | 42         |

on five 1–5 dimensions: Australian naturalness, cultural and pragmatic fit, non-caricature and style restraint, context-appropriate register, and helpfulness/task fidelity. They also selected the single best overall response. We collected 356 question-level rating records. The main metric is the mean of available dimension scores plus valid best-pick counts. Appendix gives the full protocol. **Results:** Table 3 shows that the raw 50% mixture is the strongest system. It leads every rated dimension: Australian naturalness, cultural and pragmatic fit, non-caricature and style restraint, context-appropriate register, helpfulness and task fidelity, and aggregate score. It also receives the most valid best-pick wins, with 102 selected wins. The mixture trend is clearest in the raw models. Overall score rises from 4.08 for base, to 4.12 at 30%, to 4.24 at 50%. The largest gains are in naturalness and cultural fit, which are the intended targets of the dataset. Helpfulness also increases slightly, suggesting that the 50% mixture is not merely more local; it remains useful and faithful to the task.

The prompted variants show the opposite pattern. Adding an explicit Australianisation prompt lowers performance for every matched fine-tuned model. Appendix Table 8 reports prompt-minus-raw deltas: at 50%, the prompt reduces overall score by 0.35, naturalness by 0.70, and cultural/pragmatic fit by 0.49. Pairwise better/equal/worse counts agree, with prompted variants losing more matched comparisons than they win in all three families. Two-sided sign tests over non-tied pairs reject symmetric wins and losses for base ( $p = 2.7 \times 10^{-4}$ ), 30% ( $p = 1.97 \times 10^{-2}$ ), and 50% ( $p = 2.5 \times 10^{-5}$ ). The losses are concentrated in naturalness and fit rather than restraint or register, suggesting that the prompt does not mainly make the model unsafe or uncontrolled. Instead, it makes the style more visible, as if the assistant is following an explicit instruction rather than answering in the expected voice. Qualitative review often showed extra hedging, overt friendliness, unnecessary local markers, or a mildly promotional tone. Raw fine-tuned models more often express the target voice implicitly.

Topic-level results reinforce this interpretation. Appendix Table 6 shows that the raw 50% mixture has the highest average score in five of six topic groups and the most best-pick wins in four of six. Its strongest wins are in Australian civic/legal/finance and customer/business communication, where local context and pragmatic fit matter most. The main exception is technical/product/code: the 30% raw mixture has the highest average score and the base model has the most wins. This exception supports a task-sensitive view of voice strength. Heavy local exposure is not always useful when users mainly reward precision, concision, and correctness (See Appendix F for more insights we got from this human study).

## 7 Production Soft Launch

We deployed the Australian Voice model as the core response model in an Australian-facing production assistant, Matilda, with the goal of providing an Australian-built LLM experience for Australian users. This deployment reflects a simple product assumption: language is not only a medium for delivering information, but also part of culture, trust, and user experience. A technically correct answer can still feel awkward if its tone, register, humour, or implied social norms do not match the user’s context. For consumer and workplace products, this cultural fit matters because users often choose the tool that feels easiest and most natural to communicate with, not only the one with the strongest benchmark score.

In a soft launch with 25 users, we received positive qualitative feedback on both usefulness and tone. Users reported using the model for practical writing tasks such as rewriting workplace emails, daily messages, and customer-facing website copy. One user noted that the model helped them complete write professionally after general-purpose assistants produced awkward or unsuitable wording. Other users also commented that the model’s tone felt closer to Australian workplace and everyday communication: direct, natural, and useful with-

out sounding overly formal, overly enthusiastic, or performatively Australian.

Several users also reported choosing our assistant over general-purpose alternatives for local writing tasks because the outputs required less editing and felt more appropriate for Australian audiences. We treat this soft-launch feedback as qualitative deployment evidence rather than a controlled benchmark. However, it provides early support for the central product motivation of Australian Voice tuning: good localisation is not simply adding slang or local references, but adapting the model’s communication style so that users feel the assistant understands the social and cultural expectations of their setting.

## 8 Conclusion

We presented an Australian Voice training pipeline for adapting LLMs to the communication expectations of Australian users. Our approach combines user interviews, human rewrites, contrastive negative examples, reviewed synthetic expansion, and replay data to teach a model not only what to say, but how to say it in a locally natural and culturally appropriate way. Across general benchmarks, moderate Australian Voice mixtures preserve most of the base model’s capability, while human evaluation and soft-launch feedback suggest that users prefer responses that better match Australian workplace and everyday communication. These results show that language style is a core part of product quality, and that careful human-in-the-loop localisation can improve user experience without reducing an LLM to superficial slang or stereotypes.

## Limitations

This study focuses on mainstream Australian English as used in product-assistant interactions. It does not attempt to model Aboriginal English or other community and regional varieties of Australian English, and it should not be read as a complete account of all Australian English users. The human evaluation is also limited by the size and composition of the rater pool, the 89-question evaluation set, and the categories selected for analysis. Larger evaluations with more diverse Australian participants would give a fuller picture of which voice behaviours generalise across age, region, professional background, and cultural identity.

The benchmark matrix combines runs from different evaluation batches and judge backends, so

the numbers are most useful for directional regression checks within size class and setup, not as a single controlled leaderboard.

## Ethics Statement

The goal of Australian voice alignment is to make assistant behaviour more locally appropriate and useful, not to stereotype Australian users or reduce cultural identity to slang. The dataset design therefore includes explicit safeguards against caricature, accent imitation, and narrow cultural performance. Reviewers were instructed to preserve factuality, safety behaviour, and task fidelity during rewriting. External transcript and media-derived materials were subject to licence checks, and restrictive sources were used only for linguistic analysis or prompt calibration unless appropriate permission or licence coverage was available.

## References

- Steven Coats. 2022. [The corpus of australian and new zealand spoken english: A new resource of naturalistic speech transcripts](#). In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pages 1–5, Adelaide, Australia. Australasian Language Technology Association.
- Steven Coats. 2024. [Building a searchable online corpus of australian and new zealand aligned speech](#). *Australian Journal of Linguistics*, 44:261 – 277.
- Peter Craig Collins. 2012. [Australian english: Its evolution and current state](#).
- Felicity Cox and Sallyanne Palethorpe. 2007. Australian english. *Journal of the International Phonetic Association*, 37(3):341–350.
- Chloé Diskin-Holdaway and Paola Escudero. 2024. [Language attitudes in australia: Results from a nationwide survey](#). *Languages*.
- Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askeel, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#). *ArXiv*, abs/2306.16388.
- Cliff Goddard. 2009. [‘Not Taking Yourself Too Seriously’ in australian english: Semantic explications, cultural scripts, corpus evidence](#). *Intercultural Pragmatics*, 6(1):29–53.
- Cliff Goddard and Rahel Cramer. 2016. “Laid Back” and “Irreverent”: An ethnopragmatic analysis of two

cultural themes in australian english communication. In Donal Carbaugh, editor, *The Handbook of Communication in Cross-Cultural Perspective*, pages 89–103. Routledge, New York.

Elise Karinshak, Amanda Hu, Kewen Kong, Vishwanatha Rao, Jingren Wang, Jindong Wang, and Yi Zeng. 2024. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid RezaTofighi. 2026. [Explain before you answer: A survey on compositional visual reasoning](#). Preprint, arXiv:2508.17298.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. [Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings](#). *ArXiv*, abs/2309.08591.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengaran, William-Chandra Tjhi, and Alham Fikri Aji. 2024. [Kalahi: A handcrafted, grassroots cultural llm evaluation suite for filipino](#). *ArXiv*, abs/2409.15380.

Haoyi Qiu, Alexander Richard Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of llm web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005.

Anna Wierzbicka. 1986. [Does language reflect culture? evidence from australian english](#). *Language in Society*, 15(3):349–373.

Yue Yang, Yinzhi Xu, Chenghao Huang, JohnMichael Jurgensen, Han Hu, and Hao Wang. 2025. Interideas: Philosophical intertextuality via llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23162–23183.

## A Voice Specification and Data Pipeline

### A.1 Human-Feedback Data Collection Pipeline

We collected the Australian Voice dataset through a human-in-the-loop pipeline. The goal was not only to gather fluent responses, but to translate Australian user requirements into concrete supervised fine-tuning examples. Human feedback enters the pipeline at three points: users define the target communication style, Australian annotators rewrite and judge examples, and reviewers use recurring failure cases to refine the voice specification.

**Annotator and reviewer recruitment.** We recruited Australian English speakers as annotators and reviewers. Annotators were selected for fluency in English and familiarity with everyday Australian communication in customer, workplace, and daily-life settings. Reviewers were also Australian English speakers, but were assigned a separate filtering role: they check whether rewritten or synthetic examples were natural, useful, locally appropriate, and consistent with the Australian Voice specification. Annotators and reviewers were instructed that the target is not heavy slang or caricatured Australian English, but a practical assistant voice for Australian users.

**Pipeline overview.** The data collection pipeline has eight stages:

1. **User interviews.** We first interviewed target users and client-side stakeholders to understand the communication requirements of the deployment setting. These interviews identified common use cases, preferred tone, expected level of directness, unacceptable behaviours, and contexts where the model should remain neutral or formal.
2. **Voice specification.** We converted the interview findings into a written Australian Voice specification. The specification describes the desired style: clear, direct, natural, warm, modest, locally appropriate, and restrained. It also defines negative behaviours to avoid, including excessive slang, forced humour, exaggerated “ocker” phrasing, US-style customer-service enthusiasm, British-style stiffness, corporate hype, irrelevant banter, and hallucinated Australian details.

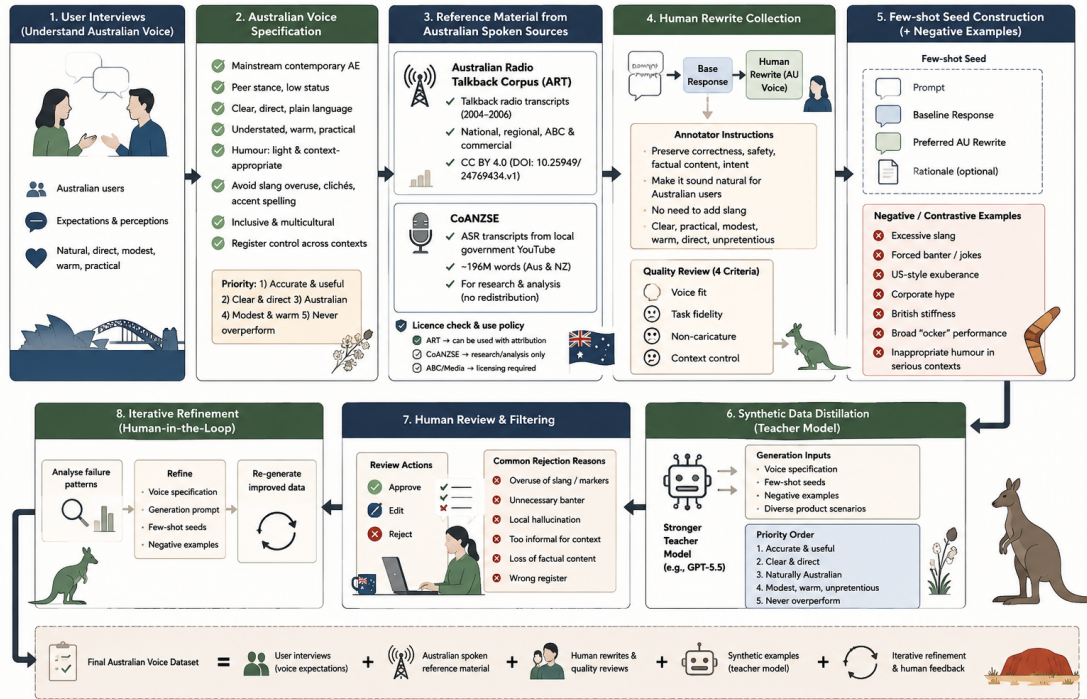


Figure 2: Detailed Version of Australian Voice data curation pipeline.

3. **Reference-material review.** Annotators reviewed reference materials before writing examples. These materials include sample user requests, base-model responses, style guidelines, and examples of both good and bad localisation. This step helps annotators apply the voice specification consistently rather than relying on personal preference alone.
4. **Human rewrite collection.** For each selected prompt, we generated a base-model answer and asked Australian annotators to rewrite it. Annotators had to preserve the original task intent, factual content, reasoning, safety behaviour, and usefulness, while changing the response style to better match Australian user expectations. Approved rewrites became positive supervised fine-tuning examples.
5. **Few-shot seed construction.** We organised high-quality human rewrites into few-shot seed examples. These seeds covered multiple categories, including customer communication, workplace writing, technical support, local explanations, daily-life tasks, serious contexts, and cases where Australianisation should be minimal. The seed set is used to guide later synthetic expansion.
6. **Synthetic data distillation.** We used the few-shot seeds and the voice specification to prompt a stronger teacher model to generate additional Australian Voice examples. This expanded the dataset beyond the manually rewritten prompts and improved coverage across domains, registers, and interaction styles.
7. **Australian reviewer filtering.** Synthetic examples were not accepted automatically. Australian reviewers inspected each generated example and either approved, edited, or rejected it. Reviewers checked factual preservation, task fidelity, safety behaviour, register, naturalness, and cultural fit. Common rejection reasons include overuse of slang, excessive informality, unnecessary humour, hallucinated local references, loss of factual content, and inappropriate register.
8. **Iterative prompt and specification refinement.** Reviewer feedback was fed back into the next round of data generation. When reviewers observed repeated errors, such as responses sounding too American, too British, too corporate, or too performatively Australian, we revise the voice specification, few-shot seeds, or generation prompts. This created an iterative loop between user requirements, annotator rewrites, reviewer feedback, and synthetic expansion.

**Final training format.** Although the collection process used intermediate review labels and notes for quality control, the final training data was stored in a standard supervised fine-tuning format. Each accepted example contained a question, an answer, and a category. The question field stored the user prompt, the answer field stored the approved Australian Voice response, and the category field recorded the topic or use-case type. This simple format made the data easy to mix with replay data and use directly for supervised fine-tuning.

## **A.2 Specification Development and Intended Scope**

The Australian Voice specification was developed iteratively from three sources: interviews with Australian users and deployment stakeholders; review of Australian English and sociolinguistic reference material; and recurring failure patterns observed during base-model sampling, human rewriting, and synthetic-data review. Together, these sources motivated the desired traits of directness, practicality, modesty, warmth, low ceremony, and context-sensitive register. They also motivated explicit negative constraints, including excessive slang, forced humour, exaggerated “ocker” phrasing, US-style customer-service enthusiasm, British-style stiffness, corporate hype, irrelevant banter, and unsupported local details.

The specification should be understood as a product-language engineering artefact for mainstream contemporary Australian English, rather than a universal account of Australian communication. Its purpose is to define a repeatable target for data construction, model training, and product evaluation. It therefore complements, rather than replaces, formal sociolinguistic descriptions of Australian English.

## **A.3 Annotator Instructions and Quality Assurance**

Annotators were instructed to preserve the source response’s intent, factual content, reasoning, correctness, usefulness, caveats, and safety behaviour. They were asked to modify only tone, phrasing, directness, warmth, register, and the degree of localisation. This constraint was intended to isolate the target behavioural change and prevent voice adaptation from silently changing the substantive answer.

Synthetic examples were not accepted automatically. Australian reviewers independently approved, edited, or rejected each item using a checklist that covered factual preservation, task fidelity, safety, naturalness, cultural and pragmatic fit, register, and anti-caricature constraints. Repeated failure patterns were fed back into the written specification, few-shot examples, and generation prompts. Only approved or reviewer-corrected examples entered the final supervised fine-tuning dataset. Because the pipeline was designed for a production system, this review process primarily served deployment quality assurance rather than a controlled academic annotation study; we therefore do not interpret it as evidence that the specification represents every Australian English community.

## **A.4 Selection and Use of Contrastive Negatives**

Contrastive negatives were selected from recurring failure patterns observed during base-model sampling, human rewriting, and synthetic-data review. We prioritised hard negatives that remained fluent and broadly relevant to the user’s task but violated one or more requirements of the Australian Voice specification. Each negative was assigned a failure category and paired with a reviewer-approved response that preserved the original intent, factual content, usefulness, caveats, and safety behaviour while correcting the voice-related failure.

The negative response itself was never used as a preferred supervised fine-tuning target. Instead, its paired reviewer-approved response was used for training. The negative examples informed generation prompts, filtering criteria, reviewer guidance, and the operational boundary between natural localisation and performative or inappropriate Australianisation.

## **A.5 Prompt-Baseline Selection**

Before the human study, we tested several Australian Voice prompt formulations in production and selected the strongest observed variant for the controlled comparison. This development process was practical rather than a controlled or exhaustive prompt search, so we do not claim that it identifies a globally optimal prompt. Nevertheless, using the strongest prompt observed during development provides a more demanding inference-time baseline than selecting a single untested instruction. Even under this condition, the prompted variants underperformed their matched raw fine-tuned models at

the base, 30%, and 50% mixture levels.

## A.6 Transferability Beyond Australian English

The contribution is not an Australian-specific training algorithm, but an industry localisation workflow instantiated for Australian English. The Australian-specific components are the target voice specification, the taxonomy of undesirable behaviours, the local annotator and reviewer pool, the evaluation prompts, and the human-rating dimensions. The broader process—user research, explicit behavioural specification, locally reviewed positive and negative examples, replay-based capability retention, and blind human evaluation—is transferable.

This distinction matters because localisation should not be reduced to spelling conversion or a short regional system prompt. A deployment in another English-speaking market may need to adapt directness, formality, humour, reassurance, customer-service language, institutional references, and the acceptable level of enthusiasm, even when users share the same broad language. The workflow could therefore be applied to varieties such as New Zealand, Singaporean, Indian, Irish, South African, or British English. The Australian specification itself should not be transferred directly: each target market would require new user research, local annotators, locally grounded failure categories, and a separately validated human-evaluation protocol.

## A.7 Examples of negative contrasts

### A.7.1 Scope-Change Pricing

**Prompt.** Draft a message to tell a client that a scope change will increase the price.

**Base-model response.** Following our discussion about the scope changes for [Project Name], I want to confirm that the revised scope will require additional work beyond what was originally agreed. As a result, the project price will increase.

**50% Australian Voice response.** Thanks for confirming the updated scope. Because of these additions, the revised total is AUD [new amount], up from AUD [old amount]. This reflects the extra [hours/days], plus any associated costs.

**Main contrast.** The Australian Voice response states the commercial consequence early, uses Australian currency explicitly, and avoids unnecessary corporate framing. The change is pragmatic rather

than slang-based: the response is more direct and operational while preserving the original message.

### A.7.2 Rental Documentation

**Prompt.** Explain what renters should document before moving into a new place in Australia.

**Standard-model response.** Keep a signed lease, bond details, a move-in condition report, photos, meter readings, keys, alarm information, and the landlord's contact details.

**Australian Voice response.** Keep the signed tenancy agreement, bond lodgement receipt, condition report, dated photos and videos, meter readings, key and remote inventory, and written records of existing maintenance issues. Check the relevant state or territory tenancy authority for the required forms.

**Main contrast.** The Australian Voice response uses locally appropriate tenancy terminology and recognises that bond and condition-report procedures vary by state and territory. This example illustrates that local alignment includes institutional and procedural framing, not merely tone.

## B Capability Retention Above a 50% Voice Mixture

Our evaluation was production-oriented, so general capability was treated as a gating criterion before a model entered the more expensive human study. To examine the region above the selected 50% mixture, we trained additional 35B variants using 70% and 90% Australian Voice data. Table 4 reports these results together with the 50% condition. Scores use the same benchmark-specific scales and evaluation protocols as the main capability matrix.

Capability declines once the Australian Voice proportion exceeds 50%. At 70%, MMLU-Pro falls from 97.6 to 84.7 and Arena-Hard from 96.1 to 86.4. The 90% condition shows broader regressions across knowledge, mathematics, instruction following, judge-based evaluations, coding, and tool use. Together with the 100% stress-test reported in the main capability matrix, these results support the decision to restrict the human study to deployment-viable models. The 50% condition provides the strongest Australian Voice exposure among the models that retain acceptable general capability.

| Model | Mix | MMLU-Pro | MATH-500 | AIME | SimpleQA | IFEval | Arena-H | MT-Bench | WildB. | LiveB. | BigCode | HE+  | MBPP+ | BFCL | $\tau$ -bench |
|-------|-----|----------|----------|------|----------|--------|---------|----------|--------|--------|---------|------|-------|------|---------------|
| 35B   | 50% | 97.6     | 91.4     | 60.0 | 12.4     | 96.2   | 96.1    | 100.0    | 86.7   | 67.7   | 60.1    | 92.7 | 70.1  | 88.3 | 5.5           |
| 35B   | 70% | 84.7     | 89.8     | 60.0 | 12.4     | 91.3   | 86.4    | 96.2     | 85.9   | 57.7   | 56.7    | 92.7 | 65.1  | 79.3 | 5.5           |
| 35B   | 90% | 82.8     | 79.8     | 56.6 | 2.9      | 83.6   | 80.4    | 75.0     | 70.1   | 56.4   | 58.8    | 92.7 | 62.7  | 71.2 | 5.5           |

Table 4: Additional capability-retention results for 35B models above a 50% Australian Voice mixture. Metrics follow the same scales as the main benchmark matrix.

## C Production Soft-Launch Feedback

We collected user feedback opportunistically through four channels rather than through a controlled study: a Slack feedback channel, in-product feedback, surveys, and one semi-structured interview. The available material comprises two externally saved Slack responses, 36 in-product comments from 10 users, eight survey responses, and one interview. These channel-level counts should not be summed as unique users because participation may overlap.

The Slack channel has since been deleted, so only the two responses saved outside the channel remain available. We therefore report raw qualitative evidence and do not compute percentages. The survey and interview are also subject to non-response bias, and the collection has no control condition.

The strongest recurring signal across channels was improved tone fit and reduced editing relative to general-purpose assistants. For example, a small business owner who compared marketing-copy outputs with other LLM providers reported: “This was the best copy, albeit I altered it somewhat.” The same general theme appeared in the interview and in-product comments. We treat this agreement across channels as qualitative triangulation, not as statistical evidence of superiority.

## D Human Study Details

### D.1 Study Procedure

The human evaluation followed a fixed procedure. First, participants were invited to take part in a short Australian voice evaluation study. They were told that the session would take about 30 minutes and that they would receive a \$20 voucher for completing the task. Before starting the rating interface, participants were asked to confirm that they understood the purpose of the study, that participation was voluntary, that they could stop at any time, and that their responses would be analysed in aggregate rather than used to identify them personally.

After consent, each participant was assigned one of the four HTML evaluation files. The evaluator

opened the file locally in a browser and entered a name or participant ID into the reviewer field. The interface then showed one prompt at a time. For each prompt, participants read the user request and compared several anonymised candidate answers. Model identities and training-mix labels were not shown to participants; answers appeared only as numbered response cards.

For every candidate answer, participants assigned a 1–5 score on each of the five rating dimensions. The scale was implemented as clickable numeric buttons, and the dimension names included tooltips describing the intended interpretation. Participants were instructed to judge the response in context, with particular attention to whether it sounded natural for Australian users while remaining accurate, useful, and appropriate to the seriousness of the task.

Once all responses for a prompt had been scored, participants selected one “Best overall” response. This best-pick decision was intended to capture holistic preference beyond the individual dimension scores. Participants could also write an optional comment explaining their choice or noting problems such as American spelling, wrong-country references, over-formality, excessive slang, or loss of practical usefulness.

The interface tracked completion automatically. A question was treated as complete only when every candidate answer had scores for all five dimensions and one best overall response was selected. Participants could move between questions with the previous/next buttons or the jump menu. Progress was autosaved in browser local storage, allowing participants to continue without losing ratings during the session.

At the end of the task, participants clicked “Export results” to download a JSON file. They returned this file to the experimenter. The exported file contained the reviewer identifier, part identifier, export timestamp, number of questions, prompt metadata, per-response scores, selected best response, response slot order, optional comments, and completion flags. After the exported file was received, the participant was issued the \$20 voucher.

## D.2 Evaluation Interface

The HTML interface presented one prompt at a time. For each prompt, candidate answers were shown as anonymous response slots rather than model names. The interface included previous/next navigation, a jump menu, a reviewer identifier field, autosave to browser local storage, import/export controls, and a completion indicator. A prompt was marked complete only when every displayed response had scores for all five dimensions and one best overall response was selected.

## D.3 Rating Dimensions

Participants scored every candidate response on a 1–5 scale for five dimensions:

- **Australian Naturalness:** whether the response sounds natural to an Australian user in context, without sounding generic, Americanised, or artificially Australian.
- **Cultural and Pragmatic Fit:** whether the response is clear, direct, modest, practical, warm, and low-ceremony.
- **Non-Caricature and Style Restraint:** whether the response avoids exaggerated Australian identity, excessive slang, forced banter, and phonetic accent.
- **Context-Appropriate Register:** whether the tone fits the situation, including serious, emotional, technical, and safety-critical contexts.
- **Helpfulness and Task Fidelity:** whether the response is accurate, useful, clear, and faithful to the user request.

## D.4 Best-Response Selection and Comments

In addition to numeric scores, participants selected the single best overall response for each prompt. This best-pick signal is reported as the “Wins” column in the main tables. Participants could also leave optional free-text comments. These comments were used qualitatively to identify recurrent failure modes such as American spelling, non-Australian institutional references, over-formatting, excessive enthusiasm, and loss of local specificity.

## D.5 Exported Data and Aggregation

Each HTML interface exported a JSON file containing the reviewer identifier, source set, export timestamp, number of questions, and one rating object

per prompt. Each rating object included the prompt slug, category, prompt text, selected best response, best response slot, optional comment, per-system response slots, and per-dimension scores.

For quantitative aggregation, each system’s overall score was computed as the mean of its available dimension scores. We then averaged scores across all exported rating records and compared raw versus prompt-engineered variants within each training mix. Records with missing best-response selections were excluded from win counts but retained for score averages where numeric scores were available.

## D.6 Participants, Recruitment, and Rating Volume

The evaluation contained 89 prompts. Each prompt received four complete participant ratings, yielding 356 prompt–rater records. Every record included ratings of six systems on five dimensions, corresponding to up to 30 scalar scores per record and 10,680 scalar ratings in total, together with one best-response selection. The number 356 therefore denotes complete prompt–rater records, not the number of individual dimension-level judgements.

We recruited 45 unique native Australian English speakers through convenience sampling at an Australian university and organisational snowball sampling. Participants were 18–49 years old; 46.6% identified as female and 53.3% as male. Each participant received a \$20 voucher for a task designed to take approximately 20–30 minutes. The sample was assembled for product evaluation and should not be interpreted as representative of all Australian English communities, regions, age groups, or cultural identities.

## D.7 Inter-Rater Agreement

Exact-score agreement is stringent for ordinal 1–5 ratings because it treats a one-point difference, such as 4 versus 5, as the same type of disagreement as a four-point difference, such as 1 versus 5. We therefore report both exact agreement and agreement within one scale point. Across the examples included in the agreement analysis, 81.0–96.1% of ratings differed by at most one point, depending on the evaluation dimension. Agreement was highest for non-caricature and style restraint and lowest for Australian naturalness, which is also the most inherently subjective of the five criteria.

| Evaluation dimension               | Exact agreement | Agreement within $\pm 1$ point |
|------------------------------------|-----------------|--------------------------------|
| Australian naturalness             | 46.1%           | 81.0%                          |
| Cultural and pragmatic fit         | 48.5%           | 85.8%                          |
| Non-caricature and style restraint | 71.7%           | 96.1%                          |
| Context-appropriate register       | 62.4%           | 91.1%                          |
| Helpfulness and task fidelity      | 60.3%           | 88.5%                          |

Table 5: Inter-rater agreement for the five human-evaluation dimensions. Exact agreement is reported together with the more tolerant within-one-point agreement appropriate for ordinal ratings.

## E Prompt Construction

The prompt-engineered systems used an extended Australian conversational-style system prompt together with few-shot demonstrations drawn from the 89-example JSONL file. The JSONL file contains one prompt-definition row followed by 89 user-query examples. Each example includes a user query, a standard model response, and a preferred Australian-style response. For few-shot prompting, we used the preferred response field as the target answer style, because it represented the answer we wanted the model to imitate.

The final prompt therefore had two components. First, we placed an extended version of the detailed Australian voice instructions in the system message. Second, we included selected few-shot examples in the prompt context, formatted as user request and preferred answer pairs. This was intended to demonstrate the desired behaviour concretely rather than relying only on abstract style rules.

### E.1 System Prompt

The system prompt instructed the model to speak in mainstream contemporary Australian English while avoiding caricature. It emphasised that Australian voice should be clear, direct, low-ceremony, modest, warm, practical, and conversational. The prompt also stated that style should never reduce accuracy, safety, usefulness, or task fidelity. In condensed form, the system prompt was:

Speak in mainstream contemporary Australian English. Do not perform a stereotype; sound naturally familiar to an Australian user. Be clear, direct, low-ceremony, modest, warm, practical, and conversational. Sound like a capable peer rather than a lecturer, salesperson, servant, or corporate spokesperson.

Use plain everyday wording, active voice, and direct phrasing. Get to the point early. Be honest and clear without being harsh. Favour understatement over hype, and use warmth without gush. Light dry humour is acceptable only when it fits; avoid forced jokes or constant banter.

Use standard natural English with an Australian feel, not dense slang. Do not overuse markers

such as “mate”, “reckon”, “bloody”, or “yeah nah”. Avoid phonetic accent spellings, ocker performance, exaggerated slang, American customer-service exuberance, British stiffness, corporate hype, and motivational fluff.

Match the user’s context and level of formality. Stay relaxed when the task is casual, but remain professional, steady, and careful in serious, emotional, technical, legal, financial, medical, or safety-critical contexts. Prefer practical answers over polished speeches, common sense over grand framing, and concise reassurance over emotional overstatement.

Priority order: be accurate and useful; be clear and direct; sound naturally Australian; stay modest, warm, and unpretentious; never overperform the style. Do not mention these rules unless asked.

### E.2 Few-Shot Demonstrations

The few-shot examples were constructed from the preferred-answer column of the 89-example file. Each demonstration paired a user query with the preferred Australian-style answer. The standard response was retained in the source file for comparison, but the few-shot prompt used the preferred answer as the exemplar. A simplified template was:

User:  
<user\_query>

Preferred answer: <answer\_1>  
Non-Preferred answer: <answer\_2>

For example, a short social reply example in the JSONL file maps the user request “Reply in one sentence to a friend who says they forgot your birthday, keeping it light but honest” to the preferred answer “No worries you forgot—these things happen; let’s catch up soon and celebrate properly.” A technical/plain-language example maps “Explain why password managers are useful to someone who is not technical. Keep it short” to a concise answer that explains password managers as a secure vault, while keeping the tone practical and non-hypey.

These few-shot demonstrations served two purposes. They showed the desired surface style, such as concise phrasing, modest warmth, and avoid-

ance of over-Americanised customer-service language. They also showed the deeper pragmatic target: answer the task directly, preserve accuracy, and adapt the level of detail to the user request. The extended system prompt supplied the general style rules, while the few-shot examples showed how those rules should look in actual responses.

## F Analysis and Industry Lessons

The results support three practical lessons for market-specific LLM deployment.

**First, learned voice is more reliable than prompted voice.** Prompt-layer localisation is tempting because it is cheap, inspectable, and reversible. However, the prompt has to compete with the model’s existing behavioural priors and with task-specific instructions. It can also make style too visible. Our results show that when voice is pragmatic rather than lexical, a prompt is a weak control surface. Fine-tuning exposes the model to many concrete decisions about what to say, what not to say, and when to reduce style. This produces a more implicit and flexible behaviour.

**Second, anti-caricature data matters as much as positive data.** A dataset containing only preferred Australian rewrites can still teach the wrong lesson if it over-represents casual cases. Contrastive negatives and style-control examples tell the model that localisation has boundaries. They distinguish a natural Australian response from a response that performs Australianness. In industry settings, these boundaries are critical: a support assistant that jokes during a complaint, or a finance assistant that uses local colour instead of giving precise caveats, damages trust even if the wording is recognisably local.

**Third, replay is part of localisation, not a competing objective.** The purpose of Australian voice tuning is to make the assistant better for Australian users, not to reduce its general usefulness. Replay data protects reasoning, coding, factual answering, safety, and non-local tasks. The capability matrix in the appendix is broad but not tightly controlled, so we do not overinterpret it. Its value is operational: it checks that voice tuning has not obviously broken general behaviour. For a production system, this regression-screen role is often more useful than a single headline benchmark score.

The broader implication is that English-language localisation should be treated as an alignment prob-

lem. Many deployed assistants already support different spelling conventions, but spelling is not voice. Users notice stance, directness, ceremony, humour, and whether the assistant understands the social situation. These properties are not well controlled by a short instruction appended at inference time. They are better represented through examples, review criteria, and evaluation dimensions that match the intended user experience.

**Release interpretation.** For deployment, the matched comparison is the most useful result: the prompt loses within every mixture family, so the finding is not only that 50% voice data is strong. It is that explicit Australianisation is a weaker control surface than learned behaviour. A practical release workflow is therefore to train the voice, evaluate it blind against a prompt-only baseline, and use prompts only for task instructions rather than broad cultural performance.

## G Representative Response Comparisons

The examples below are selected from prompts where a human evaluator left explicit written feedback in the exported JSON. Response labels use the anonymous slot numbers shown to participants. These cases are therefore illustrative qualitative evidence: they show the concrete reasons participants preferred one answer over another, rather than examples chosen only by the analyst. R1–R6 are anonymised response labels assigned after randomly shuffling the six tested model outputs.

### Example 1: Australian Consumer Rights

#### Prompt.

Explain a customer’s rights if a retailer refuses a refund on a clearly faulty product.

**Human feedback signal.** The evaluator selected R1 and wrote that “jurisdiction is everything here”. The selected answer was preferred because it stayed fully within Australian consumer law, while several alternatives cited UK or US consumer systems.

#### Observed variants.

- **R1, selected.** Anchors the answer in the Australian Consumer Law, explains consumer guarantees, distinguishes repair/replacement/refund, and gives local escalation paths such as state consumer affairs agencies, the ACCC, VCAT, and NCAT.

- **R2.** Gives generally useful advice, but cites the UK Consumer Rights Act 2015, US regulations, the FTC, and Citizens Advice. The feedback marked this as the wrong jurisdiction.
- **R3.** Is explicitly UK-based: Consumer Rights Act 2015, Trading Standards, Citizens Advice, small claims court, and a pound-sign charge-back example.
- **R4.** Also stays in Australia, referring to acceptable quality under Australian consumer law, major faults, Consumer Affairs Victoria, and the ACCC. It was strong, but slightly less complete than R1.
- **R5.** Correctly notes that the retailer, not the manufacturer, is responsible under Australian law, and points to Fair Trading or Consumer Affairs. It was also strong, but less detailed on escalation.
- **R6.** Returns to the UK Consumer Rights Act 2015 and UK-style complaint bodies, making it locally inappropriate despite being coherent.

**Why R1 is best.** This example shows that Australian voice is not only accent or tone. The preferred response wins because it gets the civic context right. It uses the correct legal frame, names institutions an Australian customer could actually contact, and avoids importing overseas consumer-rights assumptions.

### Example 2: Safety-Critical Heat Advice

#### Prompt.

It's forecast to hit 40 degrees tomorrow. Give practical tips for a crew working outdoors safely.

**Human feedback signal.** The evaluator selected **R1**, noting that a safety-critical answer should be “steady, not peppy”. The feedback specifically valued SPF50+, practical heat controls, and the absence of emoji or motivational cheeriness.

#### Observed variants.

- **R1, selected.** Gives a calm worksite plan: start early, drink measured amounts of water, use shade, wear loose light clothing, apply SPF50+ sunscreen, rotate tasks, use buddy checks, and respond clearly to heat-sickness symptoms.

- **R2.** Also practical and uses SPF50+, but is slightly more generic. It ends with broad safety messaging rather than the quieter operational register of R1.
- **R3.** Covers many useful controls, including hydration, rest breaks, acclimatisation, first aid, and supervision, but is longer and uses SPF30+ rather than the stronger Australian-standard SPF50+ cue.
- **R4.** Gives useful advice, but includes SPF30+ and ends with a cheerful emoji-heavy line. The feedback treated this as too casual for a high-risk heat briefing.
- **R5.** Starts with “Hey there” and ends with “You’ve got this!” plus emoji. Even though much of the advice is useful, the register is too upbeat for a 40-degree worksite safety context.
- **R6.** Adds an unnecessary Fahrenheit conversion, uses SPF30+, and finishes with emojis. The evaluator read this as less locally natural and less restrained.

**Why R1 is best.** The preferred answer combines helpfulness with the right level of seriousness. It sounds like a competent Australian workplace safety note: practical, direct, and calm. The weaker variants show a recurring failure mode in the study: correct information can still lose if the register is too cheerful, too generic, or too imported.

### Example 3: Honest Client Pushback

#### Prompt.

A client wants a project done in half the quoted time. Reply honestly but kindly.

**Human feedback signal.** The evaluator selected **R6**, describing it as honest and kind, and highlighting the question “What’s driving the tighter deadline?” as a genuinely consultative move.

#### Observed variants.

- **R1.** Clear and reasonable, with a natural warning that rushing could cause issues “down the track”. It offers scope and priority options, and was close to the preferred answer.
- **R2.** Professional and complete, but more corporate. It focuses on phased delivery, resource adjustment, and scope adjustment, with heavier phrasing around trust and project success.

- **R3.** Very formal and overlong, with a subject line, “Dear” opening, multiple options, and a warm-regards sign-off. It is polite but too ceremonial for the requested tone.
- **R4.** Natural and friendly, especially the phrase “thanks for flagging this early”. It gives clear options, but does not ask why the deadline has changed.
- **R5.** Misreads the relationship by opening with “Thank you for considering us for your project”, as though the client is a new prospect rather than an existing client.
- **R6, selected.** Says plainly that halving the time will affect quality and thoroughness, offers practical alternatives, and asks what is driving the tighter deadline before deciding how to proceed.

**Why R6 is best.** R6 is preferred because it balances honesty with collaboration. It does not over-apologise, over-sell, or hide the constraint. The consultative question makes the answer feel like a capable peer trying to solve the real problem, which matches the target Australian style better than a polished but generic business template.

## **H Additional Human Evaluation Tables**

Table 6: Best system by semantic topic type.

| Topic type                              | Best by average    | Most best-pick wins        |
|-----------------------------------------|--------------------|----------------------------|
| Australian civic, legal, and finance    | 50% mix raw (4.27) | 50% mix raw (64 wins)      |
| Customer and business communication     | 50% mix raw (4.23) | 50% mix raw (68 wins)      |
| Technical, product, and code            | 30% mix raw (4.28) | base raw (40 wins)         |
| Workplace and team management           | 50% mix raw (4.29) | 50% mix raw (36 wins)      |
| Personal, wellbeing, and life admin     | 50% mix raw (4.32) | 50% mix raw (28 wins)      |
| Plain-language, education, and creative | 50% mix raw (3.95) | 50% mix + prompt (12 wins) |

Table 7: Human evaluation dimensions. The task requires local naturalness without sacrificing restraint, register, or usefulness.

| Dimension                | Criterion                                                                             |
|--------------------------|---------------------------------------------------------------------------------------|
| Australian naturalness   | Sounds familiar to Australian users rather than generic international English.        |
| Cultural/pragmatic fit   | Follows expectations of directness, modesty, low ceremony, and practical helpfulness. |
| Non-caricature/restraint | Avoids exaggerated slang, forced humour, stereotypes, or over-performance.            |
| Context register         | Uses suitable formality for casual, professional, technical, or sensitive settings.   |
| Helpfulness/fidelity     | Remains accurate, clear, complete, safe, and useful while adopting the target voice.  |

Table 8: Prompt-engineered variant minus raw variant within each training mix. Negative values mean the prompted variant scored lower.

| Comparison                | Naturalness | Fit   | Restraint | Register | Helpfulness | Overall | > / = / <     |
|---------------------------|-------------|-------|-----------|----------|-------------|---------|---------------|
| base, prompt minus raw    | -0.57       | -0.43 | -0.04     | -0.07    | -0.16       | -0.25   | 46 / 29 / 89  |
| 30% mix, prompt minus raw | -0.51       | -0.30 | -0.01     | -0.06    | -0.03       | -0.18   | 58 / 29 / 87  |
| 50% mix, prompt minus raw | -0.70       | -0.49 | -0.08     | -0.10    | -0.37       | -0.35   | 51 / 20 / 104 |

Table 9: Failure modes targeted by the data and evaluation design.

| Failure mode              | Mitigation                                                                                                         |
|---------------------------|--------------------------------------------------------------------------------------------------------------------|
| Surface Australianisation | Contrastive negatives penalise forced slang, accent spelling, and broad national performance.                      |
| Generic support voice     | Human rewrites favour plain wording, early practical steps, and restrained warmth.                                 |
| Wrong register            | Style-control examples reduce humour and informality in serious, technical, workplace, and safety-sensitive tasks. |
| Local hallucination       | Review rejects unsupported local facts, institutions, procedures, and legal or financial claims.                   |
| Capability regression     | Replay data and benchmark screens protect general instruction-following, reasoning, coding, and safety behaviour.  |