

Technical Paper No. 1, October 2026

Evaluating AI transcription of court proceedings for use in social science and legal research

Samuel Jeffs, Claire McLean and Vere Lambert-Morris



Queensland Sentencing
Advisory Council

Overview

AIM	To evaluate the accuracy and operational viability of using OpenAI's Whisper, a free, open-weights artificial intelligence (AI) transcription model, to transcribe court proceedings for use in criminological research. This project seeks to determine if locally hosted AI transcription of court audio files is a viable alternative to official court transcripts for research purposes.
METHOD	The project used 'faster-whisper' (running the 'large-v2' model) on a standard issue laptop to process audio recordings, with no data leaving the secure environment. A sample of 100 audio files relating to sentencing hearings and decisions made in the Magistrates Court of Queensland for contravention of a domestic violence order (CDVO) offences were transcribed by the AI. Human researchers then manually corrected these transcripts to create a 'ground truth' baseline. The AI's performance was evaluated quantitatively using Word Error Rate (WER) and qualitatively through a semantic error analysis to assess the nature and severity of errors relative to the research questions of interest.
RESULTS	The AI transcribed 96.5% of words correctly across the sample of 100 cases (overall WER 3.5%, median per-case WER 3.5%), with 91% of cases having a per-case WER below 10%. A semantic error analysis of 30 cases identified 606 distinct error events (10.19 per 1,000 words), the vast majority of which were minor or moderate; only 15 error events (2.5%) were rated as critical, and critical error events were observed in 30% of cases. Audio quality was the strongest driver of errors with "partially clear" recordings producing roughly twice the WER and roughly four times the critical error rate of "mostly clear" recordings. Errors were dominated by phonetic garbling and grammatical "tidy-ups". Omissions of content (where speech is dropped or replaced with benign filler) accounted for 40% of all critical errors and are the type of error least visible to a reader of the AI generated transcript.
CONCLUSION	Locally-hosted Whisper transcription is operationally feasible, cost-effective and information-security compliant to generate transcripts that can be used for research purposes. The AI-generated transcripts should be treated with caution due to the potential for critical errors of relevance to the research, but are accurate enough to function as a navigation layer over the audio such that researchers can locate and skim relevant parts of a hearing in a fraction of the time previously required, while retaining the audio as the source of truth for any substantive coding decision. On this basis, AI transcription is recommended as a research-workflow tool, not as a replacement for verified transcription.

Introduction

The Queensland Sentencing Advisory Council received Terms of Reference from the Attorney-General to review certain aspects of sentencing for domestic violence (DV) offending, including how sentencing trends and outcomes for contravention of a domestic violence order (CDVO) offences may have changed following an increase to maximum penalties in 2015.¹

A CDVO offence means a DVO condition made to protect the aggrieved from DV behaviour has been breached. The broad definition of DV,² variability in DVO conditions³ and police charging practices mean the breaching conduct varies widely and a single offence can include a range of conduct and actions.⁴ CDVO behaviour can include assaulting the aggrieved, damaging property, and making threats, as well as behaviour that would not be an offence without the presence of the DVO condition, such as contacting the aggrieved when there is a no contact condition. While any breach of a court order is serious, the nature and seriousness of the conduct is one of the crucial factors to be considered in determining a sentence that is just in all the circumstances.

To understand how sentencing practices have changed following the increase in maximum penalties, the Council needs to understand the types of conduct that resulted in a CDVO sentence, how frequently each type occurred, and what sentences were given for different behaviours.

Unfortunately, no agency in the criminal justice sector collects quantitative data on the circumstances of a person's offending in a way that enables sentencing outcomes to be analysed based on the type of conduct.

Important information about the nature of conduct involved in CDVO cases sentenced is not captured in a structured way within administrative data systems.

Furthermore, databases are structured in a way that makes it difficult to administratively link, at scale, the original DVO, and any associated conditions, to the breach itself.⁵

Assessing how the nature of the conduct sentenced may have changed over time is therefore impossible without undertaking the labour-intensive process of extracting this information from hard copy court files, police court briefs and sentencing hearing transcripts.

Compounding this challenge, is that the overwhelming majority of cases sentenced for CDVO, particularly where it is the most serious offence sentenced within the matter, are dealt with in the Magistrates Court, and these court hearings are not routinely transcribed.

Given the prohibitively high cost of obtaining official court transcripts⁶, an alternative methodology was piloted: obtaining audio recordings of the proceedings and using automatic speech recognition (ASR) to generate transcripts, which could subsequently be coded and thematically analysed for the research project.

This paper evaluates the accuracy and operational viability of using OpenAI's Whisper, a free, open-weights AI transcription model, on a sample of 100 court proceeding audio files. While existing research often evaluates ASR models on general conversational datasets, interviews or highly structured medical dictation,⁷ this project assesses AI transcription within

¹ *Criminal Law (Domestic Violence) Amendment Act 2015* (Qld) s 7 amending, *Domestic and Family Violence Protection Act 2012* (Qld) s 177 ('DFVPA').

² DFVPA (n 1) s 8.

³ Ibid pt 3, div 5; See discussion on DVO conditions in Queensland Sentencing Advisory Council, *The Tangled Web: Examining Domestic and Family Violence Sentencing Reforms* (Final Report, 2026) 96–7.

⁴ See discussion on police charging practices in Queensland Sentencing Advisory Council (n 3) 97–9.

⁵ In Queensland, there have been two attempts to 'match' a CDVO with a DVO. One involved matching criminal courts CDVO charges to originating civil courts DVO orders, with only 64% of orders able to be matched, see ARTD Consultants, *Southport Specialist Domestic and Family Violence Court - Process and Outcomes Evaluation 2017-21* (Final Report, December 2021) 117. Details of the analysis and methodology are in Appendix 1, 168–9. A second approach used an alternative methodology to match civil courts DVO data to police CDVO charge data, see Queensland Government Statistician's Office, Queensland Treasury, *Breaches of Domestic Violence Orders in Queensland, 2008–09 to 2017–18* (Crime Research Report, 2021) 3, 42.

⁶ At the time of scoping the project, transcripts could cost up to \$554 per hour of court hearing, depending on how quickly the transcript is required, as compared to audio files which cost around \$45 per hour of court hearing.

⁷ Joel Jai Wei Ng et al, 'Evaluating the Performance of Artificial Intelligence-Based Speech Recognition for Clinical Documentation: A Systematic Review' (2025) 25 *BMC Medical Informatics and Decision Making* 236, Helen Eftekhari, 'Transcribing in the Digital Age: Qualitative Research Practice Utilizing Intelligent Speech Recognition Technology' (2024) 23 *European Journal of Cardiovascular Nursing* 553.

the specific context of sentencing in the Queensland criminal justice system.

The main research question is whether local AI transcription is a viable and cost-effective method for transcribing court proceeding audio files for use in criminological research. We address this through three sub-questions:

1. How accurate is the AI model at transcribing the audio files?
2. What effects do transcription errors have on the underlying research? (i.e. How often do they change the meaning of what is being transcribed)?
3. Can AI transcription be operationalised on standard issue equipment?

Information security

Because court recordings often contain unredacted, personally identifying information and information that may be prohibited from disclosure by law or by court order, the audio files used in this project are classified as Protected. Consequently, uploading these audio files to commercial cloud-based AI services (such as the standard OpenAI API or commercial transcription web services) is prohibited.

To navigate this, the project undertook a Foundational Artificial Intelligence Risk Assessment (FAIRA). The FAIRA is a Queensland Government governance framework used to evaluate the ethical, legal, and security risks of deploying AI solutions. The FAIRA process determined that the use of AI transcription was viable provided the model was run locally on a department computer and did not transmit data to third parties. These constraints shaped the choice of transcription technology by narrowing the project to transcription systems that could be run entirely on local infrastructure.

The Whisper model

To meet these information security requirements, the project utilised Whisper, an open-weights ASR system developed by OpenAI. In this context, ‘open-weights’ means the core files required to run the AI are freely available to the public. Because the model’s architecture (the underlying software design) and weights (the numerical rules the model learned during its training) are publicly available, the research team was able to download the model and run it locally on a secure, departmentally managed computer. This ensured that no Protected data ever left the secure environment.

Whisper is an encoder-decoder transformer-based model. This means it uses a two-step process: an ‘encoder’ listens to the raw audio and turns it into a mathematical representation, and a ‘decoder’ reads that math and translates it into written text. The encoder part of the process processes the raw audio and converts it into a type of image called a log-Mel spectrogram (see Figure 1). A log-Mel spectrogram maps the frequencies of the sound over time. This encoder process essentially converts the audio into a mathematical representation.

The decoder part functions similarly to a large language model by taking the mathematical representation of the audio as an input, as well as the previous words transcribed as a second input, and predicts the most mathematically probable sequence of words.

Figure 1: A log-Mel spectrogram

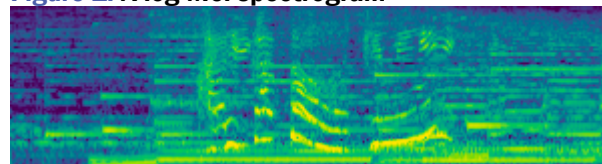


Image cropped from Alec Radford et al, 'Robust Speech Recognition via Large-Scale Weak Supervision' (2023) 202 Proceedings of Machine Learning Research 28492, Figure 1
<<https://proceedings.mlr.press/v202/radford23a.html>>.

This means that Whisper is trying to predict text that balances what it hears (based on the mathematical representation of the audio) and what makes logical sense as a sentence (in the same way that large language models predict words based on mathematical probability). The decoder uses a ‘beam search’ algorithm to explore multiple possible sentence structures simultaneously. It selects the sentence that offers the best compromise between matching the audio and forming a coherent sentence given the context provided.

The Whisper model is trained on 680,000 hours of multilingual audio collected from the internet. While highly capable, relying on internet-derived training data introduces known biases. For example, the model has been observed attempting to self-censor profanity or sensitive anatomical terms by substituting them with similar-sounding benign words. Understanding how these biases and limitations manifest when transcribing courtroom proceedings is a primary focus of this evaluation.

OpenAI released Whisper in several sizes: tiny, base, small, medium, and large. Smaller models are very fast and can run on slower computers, but they are less accurate. Larger models require more computer memory and take longer to process, but they are significantly more accurate because they have learned

more complex language patterns. Given the difficult nature of courtroom audio we chose the 'large' model to ensure the highest possible accuracy.

Within the 'large' category, OpenAI has released multiple versions over time. The original 'large' model was updated to 'large-v2' in December 2022. This version was trained for longer and generally provides more reliable transcription. A newer 'large-v3' model was released in November 2023, but it was substantially different. It was trained on 1 million hours of 'weakly labelled' audio (where the transcripts paired with the audio might not be perfectly accurate) and 4 million hours of 'pseudo-labelled' audio (where the audio was transcribed by the older large-v2 model itself, rather than a human). While OpenAI reported that large-v3 improved performance across a wide variety of languages, initial testing by our team suggested it actually made more errors when transcribing Queensland court recordings. We did not formally measure this difference as part of this project, but it was enough to deter us from using it.

OpenAI also released a 'turbo' model in September 2024. Turbo is a variant of large-v3 designed to be much faster, but it is less accurate than the standard large models. Because this research project relies heavily on accurate transcription, we chose to use large-v2. It produced the most accurate transcripts for our specific context, even though it processes audio at a slower pace.

Finally, instead of using the standard Whisper software, we used 'faster-whisper'. This is a reimplementaion of the Whisper model that is up to four times faster than the standard model for the same accuracy, and it uses significantly less computer memory. It achieves this by applying several performance tricks such as weights quantisation (reducing the precision of the model's numbers to save memory), and layers fusion (combining multiple calculation steps into a single step to save time).

The transcription process

The audio files were obtained via the secure QTranscripts online platform, and downloaded to a local computer. We downloaded the Whisper model weights from Hugging Face, a popular online platform used by developers to host and share AI models, including open-weights models like Whisper. We also installed the faster-whisper Python package.

To run the transcription, we used a standard agency laptop equipped with 16 gigabytes of system memory (RAM). The system did not have a dedicated graphics

processing unit (GPU), meaning all processing was handled by the laptop's central processing unit (CPU). The audio files were provided to us in standard .mp3 format, which the software processed without any need for conversion.

As we relied entirely on a standard issue laptop CPU, the transcription speed was generally 2 to 3 times real-time. This means a 1-hour audio recording took between 2 and 3 hours to transcribe. This demonstrates that this work can be done on standard office equipment. If a dedicated GPU were available, processing times would be drastically reduced. For context, at the time of undertaking this research, the faster-whisper developers reported that transcribing a 13-minute audio file using the large-v2 model takes just 1 minute and 3 seconds on an NVIDIA RTX 3070 Ti graphics card, compared to 2 minutes and 23 seconds using the standard OpenAI implementation on the same hardware.⁸ In contrast, our setup took about 30 minutes to transcribe a 13-minute file.

Whisper processes audio in fixed 30-second windows. By default, the model passes the transcription from one segment forward as context for the next, helping maintain consistency (for example, spelling a person's name the same way throughout). However, if a segment's output appears unreliable (indicated by repetitive text or very low confidence), the model automatically discards the previous context and retries the segment independently, preventing errors from cascading through the transcript.

To manage silent passages, Whisper has a rudimentary voice activity detection capability built into its architecture. During training, the model learned to predict a special 'no speech' token when a segment contains only silence or background noise. Separately, the faster-whisper software offers an optional external Voice Activity Detection (VAD) pre-processing step using a different AI model called Silero VAD. When enabled, this analyses the entire audio file before transcription begins, identifying speech regions and only passing those regions to Whisper. This can improve accuracy by preventing the model from attempting to transcribe ambient noise. This external VAD filter was not enabled in our configuration, meaning the transcription relied on Whisper's own internal speech detection to handle silent passages.

When running the transcription, we explicitly set the model's target language to English. This forces the model to only predict English words and stops it from accidentally trying to translate background noise into another language. We also did not provide an initial prompt. An initial prompt is introductory text you can

⁸ SYSTRAN, 'faster-whisper: Faster Whisper Transcription with CTranslate2' (GitHub repository), <https://github.com/SYSTRAN/faster-whisper>.

give to the whisper model AI to guide its spelling or context before it starts transcribing.

Our python script was set up to output a plain text file (.txt), appending ‘_faster-whisper_large-v2’ to the file name of the original audio file. It also output timestamps and confidence scores for each segment of speech, using the following format:

```
[00:00:00.920 -> 00:00:03.120] [C:0.61] Mr Smith, you  
are charged that on the 1st of January  
[00:00:03.850 -> 00:00:06.920] [C:0.80] you  
contravened the domestic violence order.
```

A confidence score is a rating between 0 and 1 of how sure the AI is about the text it just generated. For each segment, the software provides a probability for every individual word (a number between 0 and 1, where 1 means the model is completely certain). We calculated the segment’s confidence score by taking the average of these individual word probabilities. If individual word probabilities were not available for a segment, the system fell back to converting the segment’s overall average log-probability into a standard probability. A log-probability is a compressed way of representing very small numbers; converting it back produces a value between 0 and 1 that can be interpreted as the model’s overall confidence in the segment.

Finally, it is important to note that Whisper does not perform speaker diarisation. Diarisation is the process of identifying who is speaking (for example, labelling text as ‘Speaker 1’ or ‘Speaker 2’). We did not use any additional software to separate the speakers. Instead, we relied on the context of the conversation and the voices in the audio file to manually determine whether the magistrate, prosecutor, defence, or defendant was speaking.

Method

A sample of 100 cases was randomly drawn from the broader study population (n=1,395). The random sample resulted in 59 cases selected from two distinct time periods relevant to the broader research project, namely 2014–15 (prior to the CDVO reforms) and 41 cases from 2023–24 (subsequent to the CDVO reforms).

Audio quality classification

Initial reviews revealed a significant disparity in audio quality across the sample. Some recordings were sourced from backup tape systems that used a single microphone for the entire courtroom. In these instances, speakers were often distant from the

microphone, and background noise such as shuffling papers, scraping chairs and coughing rendered portions of the audio unintelligible even to human listeners. In contrast, other recordings generally had individual microphones for each speaker and minimal background noise.

To control for this variable, we manually categorised each audio file into one of three quality classifications: mostly clear, partially clear, or poor. A recording was classified as mostly clear where the audio was substantially clear throughout. There may have been short segments where a speaker was far from a microphone, or where there was isolated paper shuffling, but these issues generally only affected a few seconds of speech. A recording was classified as partially clear where some parts of the audio were clear and others were poor. This usually occurred where one or more speakers were far from a microphone, or where for other reasons one or more speakers could not be heard clearly. A recording was classified as poor where all speakers could not be heard clearly throughout the recording. While ‘poor’ audio files likely exist in the broader project population, none of the 100 transcripts randomly drawn for this evaluation fell into this category.

This classification is strictly about the quality of the audio itself, that is, the quality of the microphone, the proximity of speakers to the microphone, and the presence or absence of background noise or static. It is not about whether a speaker speaks ‘clearly’ in the conversational sense, for example whether they mumble or have a thick accent.

Creating the ground truth dataset

To evaluate the AI’s accuracy, we manually corrected the 100 AI-generated transcripts to create a ground truth dataset. In data science, a ground truth is a perfectly accurate baseline dataset used to test how well an AI model performs. The raw AI-generated transcripts (plain text files) were imported into Microsoft Word with Track Changes enabled. We listened to the audio recordings and corrected any errors in transcription.

To facilitate cleaner qualitative analysis later in the study, corrections were made at the word or phrase level rather than the character level. For example, if the AI incorrectly transcribed ‘aggrieved’ as ‘grieved’, we deleted the entire word ‘grieved’ and inserted ‘aggrieved’, rather than simply inserting the letters ‘ag’.

Quantitative analysis (Word Error Rate)

We calculated a Word Error Rate (WER), using the 'jiwer' Python package. WER is calculated by adding the number of substitutions (words the AI got wrong), insertions (extra words the AI added), and deletions (words the AI missed). That total is then divided by the total number of words in the ground truth (human-corrected) transcript. This gives a score representing the proportion of words that were transcribed incorrectly.

A WER of 0% indicates a perfect transcription with no errors. A WER of 100% means the number of errors equals the total number of words in the reference transcript. Notably, WER can exceed 100% in cases where the AI inserts a large number of extra words that were never actually spoken, since insertions add to the error count without increasing the size of the reference transcript.

To ensure a fair and accurate like-for-like comparison, data pre-processing was applied to the AI transcripts before calculating the WER.

First, the text was normalised. The raw AI output included metadata, such as timestamps and the confidence scores discussed above (for example: [00:00:00.920 --> 00:00:03.120] [C:0.61]). These annotations were removed from the text prior to analysis to ensure we were only comparing the actual spoken words.

Second, some cases involved multiple court hearings and therefore had multiple audio recordings and multiple transcripts for the one case. For these, all hearings were concatenated into a single continuous document for the case. This ensured that the WER was calculated and reported at the case level. This approach aligns the quantitative metrics with the study's metadata (such as audio quality classifications) and provides a holistic measure of transcription accuracy for the entire sentencing process.

Finally, we had to account for indistinct audio. When the human researchers were unable to understand the audio, they inserted the tag [UNCLEAR]. Because it is unreasonable to penalise an AI for incorrectly guessing words that a human also could not discern, any audio marked as [UNCLEAR] in the ground truth was entirely removed from both the human and AI transcripts before running the WER calculation. Importantly, this approach means the reported WER reflects the model's performance on intelligible speech only, rather than on the audio file as a whole.

Qualitative analysis (Semantic Error Analysis)

While WER is a standard metric, it treats all words equally. Misspelling a defendant's surname carries the same mathematical penalty as omitting the word 'not' from a pivotal part of a transcript. A semantic error analysis was conducted on a sub-sample of 30 AI generated transcripts to evaluate the practical impact of the AI's transcription errors on the research project.

The 30 transcripts were drawn from the broader sample of 100 using a stratified random sampling approach. We stratified across the two distinct time periods to ensure even representation, selecting 15 cases from 2014–15 and 15 cases from 2023–24.

Because we expected audio quality to be one of the strongest drivers of transcription error, we sought to ensure the model's performance under different conditions was assessed. To prevent a purely random sample from excluding lower-quality recordings entirely, we aimed to purposely oversample to include at least 2 cases each from the 'partially clear' and 'poor' categories.

The 'partially clear' quota was met naturally with $n=3$ 'partially clear' cases included by random selection. However, as established earlier, because the broader sample of 100 contained no cases classified as 'poor', we were unable to oversample this category. Consequently, the semantic error analysis reflects only 'mostly clear' and 'partially clear' audio files. It remains possible that the broader study population contains recordings of poor quality, but none were drawn into the evaluation sample.

We manually examined the discrepancies between the AI-generated transcripts and the human-corrected transcripts and categorised each error using a two-step coding process. First, each error was assigned a severity rating based on how much the error impacted the transcript. Second, each error was assigned a category to describe the type of error that was made (for example, phonetic garbling, omission of content, or misspelling proper nouns).

We quantified errors as error events rather than counting the number of words affected. This means, if a cluster of incorrect words altered a single continuous concept, it was logged as one single error event. Similarly, if the AI model misspelled the defendant's surname and continued to make the same mistake through the entire transcript, this was logged as a single error event. This was done to prevent systematic mistakes from disproportionately skewing the error frequencies.

This approach is different to the quantitative WER analysis described above. The WER calculation already counts every individual word error, so a repeated misspelling of a defendant's surname would be counted dozens of times across a single transcript. That is appropriate for a raw measure of word-level accuracy. For the qualitative analysis, however, we were interested in the practical impact of errors on a researcher reading the transcript. Counting the same systematic mistake once gives a clearer picture of the distinct problems a researcher would actually encounter.

The severity rating for each error was determined based on the practical implications for the quality of the transcript and the way it would affect any data extracted from the transcript as part of the broader research project on the type of conduct involved in CDVO cases. There were three ratings, minor, moderate and critical. Minor errors included non-substantive mistakes that did not alter the factual meaning of the sentence or disrupt readability. Moderate errors included substantive mistakes that garble a sentence but would not mislead a researcher coding the data of relevance for the broader research project. Critical errors include those that fundamentally alter, reverse or obscure facts required for our research project.

Each error was also assigned one category that best described the mistake. These included categories such as phonetic garbling (where the AI mishears the audio and substituted a completely different word or phrase), grammar, syntax or lexical substitution (for changes to sentence structure, punctuation or swapping words that mean the same thing), and omission of content (where entire words, clauses, sentences or paragraphs are not transcribed).

The full definitions of coding rules and thematic examples are detailed in the codebook provided at Appendix 1.

Results

Quantitative analysis (Word Error Rate)

The WER analysis was conducted on the full sub-sample of 100 cases. The corrected ground-truth transcripts comprised 237,426 words across 29.5 hours of audio (1,771.5 minutes), of which 86 cases (86%) were classified as 'mostly clear' and 14 (14%) as 'partially clear'. No cases in the sample were classified as 'poor'. Fifty cases were corrected by Researcher A

and 50 by Researcher B. The average speech rate across the sample was 134 words per minute, consistent with typical conversational speech.

Overall accuracy

Table 1: Word Error Rate Summary Table

Measure	Value
Overall	
Number of cases	100
Total ground-truth words	237,426
Audio length (minutes)	1,771.5
Errors	
Total errors	8,307
Substitutions	4,487
Deletions	2,849
Insertions	971
By duration	
Errors per minute of audio	4.7
Overall WER	
Overall WER (pooled WER)	3.5%
Per-case WER	
Average	4.9%
Median	3.5%
Minimum	0.4%
Maximum	42.3%

Throughout this section we report two metrics. The overall WER (or 'pooled' WER) is calculated by adding up all the errors across all 100 cases and dividing by the total number of words across all 100 cases. This is the measure of how accurate the AI was across the sample as a whole, and it gives more weight to longer cases (which contribute more words). The per-case WER is calculated separately for each case and then summarised (for example, by taking the median across cases). This treats every case equally regardless of its length and is useful for describing how the AI typically performs on an individual case.

The overall WER across the sample was 3.5%, meaning that 96.5% of the spoken words across the 100 transcripts were transcribed correctly. The average per-case WER was 4.9% and the median was 3.5% (IQR 2.2% to 5.5%). Per-case WER ranged from 0.4% to 42.3%. The majority of transcripts were highly accurate: 69 of the 100 cases (69.0%) had a WER below 5.0%, 91 (91.0%) below 10%, and 98 (98.0%) below 20%. Phrased another way, the AI made an average of 4.7 errors per minute of intelligible audio.

Across the 237,426-word sample, there were 8,307 word-level errors. The majority (54.0%) were substitutions, followed by deletions (34.3%) and insertions (11.7%). The relatively high share of deletions is consistent with the omission of quiet

speakers and short interjections discussed later in this paper.

Distribution

The distribution of per-case WER was heavily concentrated with most cases clustered below 5% and a long right-hand tail running out to a single extreme value (see Figure 2)⁹.

Outliers

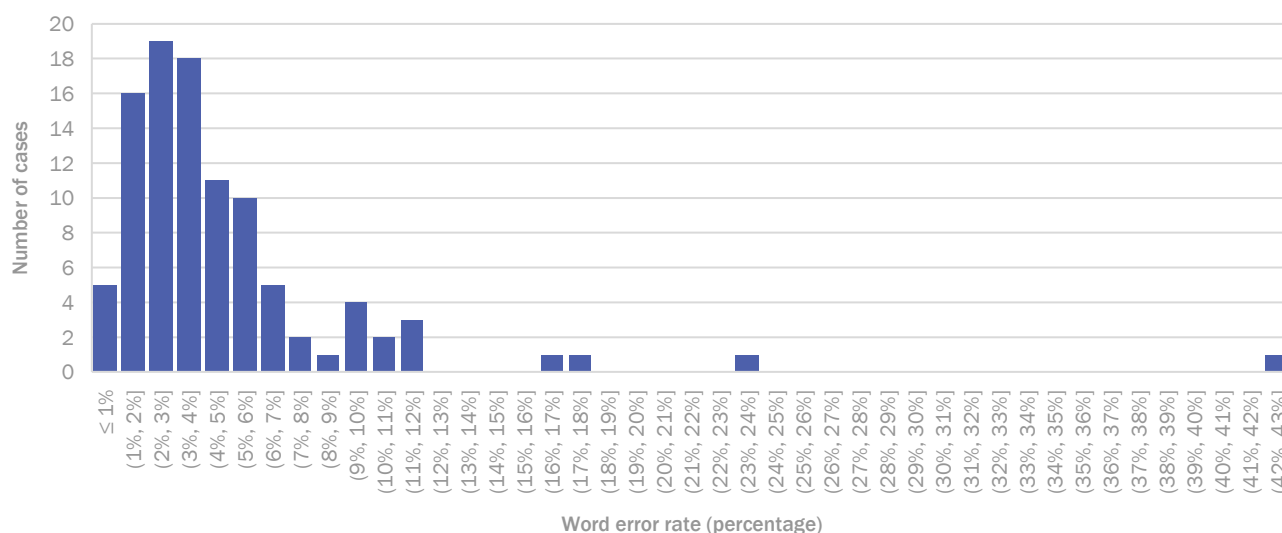
One case had a WER of 42.3%, nearly double the next-highest case (23.7%). Statistically, this case is an outlier. A standard rule (known as the Tukey upper

fence) placed the boundary for outlier values at 10.4% for our sample. This means that any case with an error rate of more than 10.4% sits outside the normal, expected range of the sample.

This case with an error rate of 42.3% had been categorised as ‘mostly clear’ and was perfectly audible to the researchers. Of the 785 words in the ground-truth transcript, 296 (37.7% of the transcript) were deletions, with entire blocks of speech completely omitted by the AI. On investigation, although the primary speaker was loud and clear, there was unintelligible background chatter in the courtroom during the segments that the AI failed to transcribe.

Overall, there were 8 outlier cases.

Figure 2: Distribution of per-case word error rates



⁹ Figure 2 shows the distribution of WER percentages. The histogram’s x-axis contains automatically assigned groupings. Borderline percentages belong in the higher category, for example a WER of 2% would be assigned to the 2-3% grouping.

Audio quality

Audio quality appeared to be the strongest driver of transcription accuracy. Records categorised as ‘partially clear’ had an overall WER more than twice that of ‘mostly clear’ recordings (6.8% vs 3.1%), with the average and median per-case WER similarly elevated (average: 6.8% vs 3.1%; median: 8.5% vs 3.3%).

Time period

There was no difference in WER across the two time periods (3.5% for both periods). The average and median WER were also similar for both time periods. This suggests that the underlying recording infrastructure produces audio of broadly similar quality across the period covered by the sample, and that the AI’s performance does not appear to be tied to the time period from which cases are drawn.

Variation between correctors

Of the sample of 100 transcripts analysed, 50 of the AI-produced transcripts were corrected by one human researcher, and 50 by a second researcher. This section looks at whether the human who corrected the transcript appears to influence the WER.

The overall WER was higher for cases corrected by Researcher B than for cases corrected by Researcher A (4.3% vs 2.8%), with the average and median per-case WER following the same pattern. It is not possible to separate genuine differences in the underlying difficulty of the cases assigned to each researcher from differences in correction sensitivity, as a more thorough corrector will identify more AI errors and so produce a higher WER. The specific rates reported should accordingly be treated as indicative rather than definitive, although the overall picture, that WER is low and concentrated in the lower single digits, holds under either researcher.

Table 2: WER by audio quality

			Word Error Rate (WER)	
Audio quality	Number of cases	Total words	Overall (pooled)	Per-case average (median)
Mostly clear	86	213,399	3.1%	4.4% (3.3%)
Partially clear	14	24,027	6.8%	8.0% (8.5%)

Table 3: WER by time period

			Word Error Rate (WER)	
Time period	Number of cases	Total words	Overall (pooled)	Per-case average (median)
2014–15	59	137,814	3.5%	4.8% (3.4%)
2023–24	41	99,612	3.5%	5.2% (3.7%)

Table 4: WER by researcher who corrected the AI generated transcript

			Word Error Rate (WER)	
Corrected by	Number of cases	Total words	Overall (pooled)	Per-case average (median)
Researcher A	50	124,114	2.8%	3.7% (2.8%)
Researcher B	50	113,312	4.3%	6.1% (4.1%)

Semantic Error Analysis

The semantic error analysis was conducted on a sub-sample of 30 cases. The transcripts of these sentencing proceedings ranged in length from 686 to 7,963 words (median 1,464 words) and collectively had a total length of 59,444 words. The sub-sample was evenly split to include 15 cases from 2014-15 and 15 cases from 2023-24. The sub-sample predominantly comprised 'mostly clear' audio (27 cases, 90%), with the remaining three classified as 'partially clear' (one from the first time period, and two from the second). The 30 cases were split evenly between two human researchers, who each manually corrected 15 of the transcripts to produce the ground truth sample.

Severity of errors

Across the 30 transcripts, 606 distinct error events were identified, equivalent to a pooled rate of 10.19 error events per 1,000 spoken words. The vast majority of errors were of low or moderate severity. Only 15 errors were rated as critical (2.5% of all errors, or 0.25 per 1,000 words), these were errors with the potential to mislead a researcher coding the substantive facts of a matter relevant to the broader research project. Critical errors occurred in 9 of the 30 cases (30.0%);

the remaining 21 cases (70.0%) contained no errors that would distort the underlying research.

Table 5: Number of error events by severity of error

Severity of error	Number	Percentage	Rate per 1,000 words
Minor	219	36.1%	3.68
Moderate	372	61.4%	6.26
Critical	15	2.5%	0.25
Total	606	100.0%	10.19

Types of errors

Errors were concentrated in a small number of themes. Phonetic garbling was by far the most common, accounting for 37.3% of all errors. The second most common type of error was grammatical changes, corrections, and lexical substitutions, making up 19.0% of errors. Hallucinations (text invented by the AI when nothing was actually being spoken) were rare (6 events, or 1.0% of errors); although it's worth noting that some 'hallucinations' were coded as 'omission of content' as described below.

To give a sense of what each theme looks like, we extracted some examples of each. These examples are illustrative rather than exhaustive and have been selected to give an indication of the types of mistakes we typically encountered.

Table 6: Number of error events by category

Category/theme	Number	Percentage	Rate per 1,000 words
Phonetic garbling	226	37.3%	3.80
Grammar, syntax, lexical substitution, or correction	115	19.0%	1.93
Domain-specific terminology	72	11.9%	1.21
Conversational filler	61	10.1%	1.03
Proper noun (name)	39	6.4%	0.66
Omission of content	37	6.1%	0.62
Proper noun (geography)	28	4.6%	0.47
Meaning reversal	22	3.6%	0.37
Hallucination	6	1.0%	0.10
Total	606	100.0%	10.19

Phonetic garbling

This was the most common type of error and covers situations where the AI substituted one or more spoken words with unrelated words that sounded similar. The result is often nonsensical or out of place in context.

AI-generated text	Human-corrected text
You have had to check it past	You have had a chequered past
This is a matter where the respondent previously indicated a clear inability to discharge .	This is a matter where the respondent previously indicated a plea of not guilty to this charge .
Probation's not an option if you go on your mistake .	Probation's not an option if you're going interstate .

Grammar, syntax, lexical substitution, or correction

These errors involve small shifts in sentence structure or word choice that do not change the underlying meaning, and frequently reflect the model 'cleaning up' the speaker's grammar.

AI-generated text	Human-corrected text
in prison	imprisonment
while it's	whilst
...and that he gave her a handwritten letter .	...and that he gave her a letter-- he gave her a handwritten letter .
she liked to help me	she liked help me

Domain-specific terminology

These errors involve the AI substituting a piece of legal, court-specific or culturally specific terminology that it appears to not recognise, with an everyday English word. We suspect that this happens for phrases that are common in court proceedings but are rare in the data on which the model was trained. Some specific phrases such as 'aggrieved' and 'SPER' (the State Penalties Enforcement Registry) were particularly challenging for the AI to transcribe correctly.

AI-generated text	Human-corrected text
bury the bail	vary the bail
crater services	Corrective Services
spur	SPER
yarding circle	yarning circles
we agreed / bereaved / McGreeve / their grave / degrees / etc	aggrieved

Conversational filler

These errors involve courtroom pleasantries, stutters, or short interjections that were either mis-transcribed or omitted, but that carry no factual weight. The AI also routinely omitted short interjections such as "Yes, your honour", "It certainly does" and "Yes, please. Ok".

AI-generated text	Human-corrected text
It does, thank you .	It does, Your Honour
I know , off you go.	Right-o , off you go.

Proper noun (name)

Names of defendants, witnesses and other parties were frequently misspelt or rendered phonetically. This type of error was very common and happened at least once in almost every case (28/30, 93.3% of cases). If a defendant's name was mis transcribed throughout the entire transcript, this was only counted as one error event, however, each name that was transcribed incorrectly was counted as a separate error event. We cannot provide examples here due to privacy considerations, although some organisation names that were mis transcribed were also included under this theme.

AI-generated text	Human-corrected text
anger care	Anglicare

Omission of content

These errors occur where a meaningful portion of speech is dropped from the transcript altogether or replaced with hallucinated text. A common pattern was the AI omitting short responses from quiet speakers, particularly the entry of a plea, for example, transcribing “How are you pleading to the charge?” without the defendant’s reply of “Guilty”. The worst examples of this error were where the AI would occasionally replace an entire paragraph worth of content with benign phrases such as repeating the phrase “Thank you, Your Honour.”

AI-generated text	Human-corrected text
He went over, and then he left	He went over, an argument ensued , and then he left
The purpose which I impose upon you today is to punish you	The purpose which I impose a sentence upon you today is to punish you
Thank you. Thank you, Your Honour. Thank you, Your Honour	There's no Queensland criminal history. That's right Your Honour, the defendant has no Queensland criminal history. Anything else you want to say about the matter?
[omission]	I do take that into account.

Proper noun (geography)

Place names were a consistent source of error. This included a range of places ranging from incorrectly transcribing street names, suburb names, and even the names of towns and regional centres.

AI-generated text	Human-corrected text
Beanley / Bingley	Beenleigh
Marlborough	Maryborough
Prime	Mackay
MARTA / Mata / morgue / Marder	Mater Hospital

Meaning reversal

These errors were uncommon. They are characterised by a small substitution that flips the direction of a fact or sentence, changing the meaning completely. The most consequential example in the sample was the first example included in the table below, as this completely reversed the outcome of a drug test.

AI-generated text	Human-corrected text
There were two analysis tests and both were clearly illicit drugs	There were two analysis tests and both were clear of illicit drug uses
She was angry.	He was angry.
No, Your Honour, his instructions aren't to contest.	No, Your Honour, his instructions are to contest.

Hallucination

This theme was used for text that was invented by the AI when nothing was actually being spoken. Note that, under the rules of our codebook, cases where the AI produced hallucinated text in place of speech that was actually present were coded as omissions rather than hallucinations (since the substantive issue is that information was lost from the transcript). These were often benign, but in one case in the sample, the AI hallucinated a specific offence date.

AI-generated text	Human-corrected text
Thank you very much.	[silence]
Thank you, Your Honour.	[silence]
There's a contravention of the... 16th of April. What's happened? Is that order still current?	There's a contravention of the... just trying to read when it was issued It must have expired mustn't it have for this new one?

Table 7: Number of errors by theme and severity

Theme	Minor	Moderate	Critical	Total
Phonetic garbling	28 (12.4%)	194 (85.8%)	4 (1.8%)	226
Grammar, syntax, lexical substitution, or correction	109 (94.8%)	6 (5.2%)	0 (0.0%)	115
Domain-specific terminology	5 (6.9%)	64 (88.9%)	3 (4.2%)	72
Conversational filler	55 (90.2%)	6 (9.8%)	0 (0.0%)	61
Proper Noun (name)	8 (20.5%)	31 (79.5%)	0 (0.0%)	39
Omission of content	6 (16.2%)	25 (67.6%)	6 (16.2%)	37
Proper Noun (geography)	3 (10.7%)	25 (89.3%)	0 (0.0%)	28
Meaning reversal	1 (4.5%)	19 (86.4%)	2 (9.1%)	22
Hallucination	4 (66.7%)	2 (33.3%)	0 (0.0%)	6
Total	219	372	15	606

Theme by severity

There were clear patterns in the severity of specific themes (see Table 7).

Grammar/syntax errors and conversational fillers were almost exclusively minor (94.8% and 90.2% respectively), reflecting the fact that they typically involve cleaning up speech dysfluency or omitting courtroom pleasantries that carry no factual weight for our research purpose.

By contrast, phonetic garbling, domain-specific terminology errors, proper noun errors, and meaning reversals were mostly moderate, reflecting their tendency to obscure or distort meaning.

Half of all critical errors were caused by omissions. Critical errors were spread across only four themes: omission of content (6 events), phonetic garbling (4 events), domain-specific terminology (3 events) and meaning reversal (2 events).

Error rates by audio quality

Audio quality had a substantial effect on error frequency. ‘Partially clear’ recordings produced more than twice the error rate of ‘mostly clear’ recordings (19.91 vs 9.44 errors per 1,000 words), and a critical error rate roughly four times higher (0.94 vs 0.20 per 1,000 words). The effect was particularly pronounced for moderate errors, which increased from 5.65 to 14.05 per 1,000 words. This is consistent with the expectation that distance from the microphone, background noise and overlapping speech create the conditions in which the AI is most likely to generate phonetically plausible but incorrect text.

These results should be interpreted with caution. Only three cases in the sub-sample were classified as ‘partially clear’, so these estimates are based on a small number of recordings (see Table 8).

Error rates by time period

Error rates were broadly similar across the two time periods. Post-reform cases had a slightly higher overall error rate than pre-reform cases (11.38 vs 9.43 per 1,000 words), driven by slightly higher rates across all three severity categories. The critical error rate was similar between periods (0.22 pre-reform vs 0.30 post-reform) (see Table 9).

Critical errors

Critical errors are an important consideration for this project. These are errors that, if undetected, would cause a researcher to miscode any of the variables we were coding as part of the broader research project.

We examined each of the 15 critical errors individually. The 15 critical errors were distributed across 9 cases. The case with the most errors accounted for 3 of the 15 critical errors.

Omission of content resulted in 6 of the 15 critical errors, in which a meaningful chunk of speech was either dropped entirely or replaced with a hallucinated ‘Thank you, your honour’. The remaining critical errors comprised 4 phonetic garbling events, 3 domain-specific terminology errors and 2 meaning reversals.

Notably, critical omission errors were sometimes mitigated by repetition elsewhere in the transcript, such as where the omitted breach behaviour was

independently described by another speaker later in the hearing. This suggests that the practical risk of miscoding a case is lower than the raw critical-error count implies, because researchers reading the transcript holistically will frequently encounter the same factual material more than once.

Variation between correctors

The semantic error analysis was conducted by a single researcher, so the error counts and categorisations are not affected by differences between people coding the errors. However, the 30 underlying transcripts were corrected by two different researchers (15 each), and this section examines whether the choice of corrector appears to influence the measured error rate. Because

each transcript was corrected by only one researcher, inter-rater reliability was not assessed.

The pooled error rate for transcripts corrected by Researcher B was higher than that of Researcher A (13.00 vs 7.76 errors per 1,000 words), as was the critical error rate (0.36 vs 0.16 per 1,000 words) (see Table 10. This difference may reflect genuine variation in the cases assigned to each coder, but it may also reflect differences in correction sensitivity.

The relatively small number of critical errors and the consistency of the broader patterns (e.g., the dominance of phonetic garbling, the rarity of hallucinations) suggest that the headline conclusions are robust to coder variation, but the specific rates reported should be treated as indicative rather than definitive.

Table 8: Rate of errors by audio quality

Audio quality	Number of cases	Total words	Rate of errors per 1,000 words			
			Minor	Moderate	Critical	Total errors
Mostly clear	27	55,174	3.59	5.65	0.20	9.44
Partially clear	3	4,270	4.92	14.05	0.94	19.91

Table 9: Rate of errors by time period

Time period	Number of cases	Total words	Rate of errors per 1,000 words			
			Minor	Moderate	Critical	Total errors
2014–15	15	36,068	3.33	5.88	0.22	9.43
2023–24	15	23,376	4.24	6.84	0.30	11.38

Table 10: Rate of errors by person who corrected the transcript

Corrected by	Number of cases	Total words	Rate of errors per 1,000 words			
			Minor	Moderate	Critical	Total errors
Researcher A	15	31,823	2.51	5.09	0.16	7.76
Researcher B	15	27,621	5.03	7.60	0.36	13.00

Discussion

This paper set out to answer three questions: how accurate is the Whisper model at transcribing Queensland Magistrates Court sentencing hearings from audio files; what practical effect do its errors have on the underlying research; and whether locally-hosted AI transcription can be operationalised on standard agency equipment.

Taken together, the WER and semantic analyses tell a reasonably consistent story, and the answers to all three questions are substantively positive, with some important qualifications around audio quality and the appropriate role of the AI output in the research workflow.

General accuracy

The main finding from the WER analysis is that the AI transcribed 96.5% of words correctly across the sample, with a median per-case WER of 3.5% and most cases (91%) having less than 10% of the words transcribed incorrectly. By the conventional benchmarks reported in the speech recognition literature, this is a high level of accuracy for unconstrained, multi-speaker, real-world audio of the kind produced by Queensland courtrooms. A 2025 systematic review of 29 studies found WER rates ranging from 0.087 in very controlled settings to over 50% in conversational multi-speaker scenarios, our court proceeding audio falls closer to the conversational end of this range as while some recurring phrases are common, each case is different.¹⁰ This was also achieved without any domain-specific fine-tuning of the model or speaker-specific adaptation. The result reflects what is available out-of-the-box from a publicly released, open-weights AI model.

That said, the distribution of errors also matters. The per-case WER ranged from 0.4% to 42.3%, with a long right-hand tail and a small number of outliers (n=8). The practical implication is that there is no guarantee that any individual transcript will be accurate; a small number of cases are likely to be substantially worse than the median, and the conditions that produce these outliers are not understood, but may be the result of overlapping speakers, background noise, or the speaker being too quiet.

Audio quality is the dominant driver of accuracy

The biggest difference in accuracy was due to the quality of the audio. Recordings classified by the researchers as 'partially clear' had roughly twice the WER of 'mostly clear' recordings (6.8% vs 3.1%), more than twice the overall semantic error rate (19.91 vs 9.44 per 1,000 words), and roughly four times the critical error rate (0.94 vs 0.20 per 1,000 words).

By contrast, the time period from which a case was drawn made effectively no difference to WER (3.5% for both 2014–15 and 2023–24) and only a small difference to the semantic error rate (9.43 vs 11.38 per 1,000 words). This was not an obvious result going in, given the decade-long gap between the two periods and the possibility that recording infrastructure had changed materially over that time. The absence of a period effect suggests that audio quality, rather than recording vintage, is what the model is responding to, and that conclusions drawn from the AI's performance on contemporary cases are likely to extend to historical recordings of comparable quality.

A summary of the findings

The two analyses are measuring different things and it is worth being explicit about how they fit together. WER captures how often the AI gets a word wrong; the semantic error analysis captures how often a transcription error has the potential to mislead a researcher coding the substantive facts of the case.

These two perspectives are largely consistent, but they are not interchangeable. The WER figure of 3.5% would, on its own, leave open the possibility that the errors would have a substantial impact on the underlying research project. The semantic analysis rules this out. Of the 606 distinct error events identified across the 30-case sub-sample, only 15 (2.5%) were rated as critical, and 70% of cases contained no critical errors at all. The errors that the model does make are dominated by phonetic garbling and grammatical "tidy-ups" that either read as obviously wrong in context (and therefore prompt the researcher to check the audio) or carry no factual weight. Hallucinations in the strict

¹⁰ Joel Jai Wei Ng et al, 'Evaluating the Performance of Artificial Intelligence-Based Speech Recognition for Clinical Documentation: A Systematic Review' (2025) 25 *BMC Medical Informatics and Decision Making* 236.

sense, where the model invented content from nothing, were rare and benign (6 events, 1.0%).

The most consequential type of error was omissions. Six of the 15 critical errors involved a meaningful chunk of speech being dropped from the transcript, often replaced by a benign phrase such as “Thank you, Your Honour”. Unlike a phonetic substitution, an omission leaves no in-text signal that anything has been lost, which makes it inherently harder for a researcher to detect from the transcript alone. This is the single most important qualification on the otherwise positive accuracy findings, and it directly motivates the workflow recommendation described below.

A useful counterweight to the omission risk is the structure of a sentencing proceeding itself for this offence. Critical factual material, particularly the conduct constituting the breach, is typically described more than once across a hearing (for example by the prosecutor and again by the magistrate in sentencing remarks). In practice we observed that omitted material was frequently recoverable from elsewhere in the same transcript, which suggests that the practical risk of miscoding a case is lower than the raw critical-error count implies.

Why these errors occur

The error patterns described above do not appear to be random. Three features of Whisper’s architecture, observed repeatedly in the sample, account for some of the errors we encountered.

One-directional processing

One weakness we observed in Whisper’s architecture was related to the way it processes transcripts from start to end. That is, when a human listens to a confusing or muffled section of audio, they might not understand it at first and mark it as [UNCLEAR]. However, as they continue to listen to the transcript, the same word, phrase or context might be spoken more clearly later in the proceeding. With this additional context, the human transcriber can go back and fix the earlier mistake.

We found that Whisper cannot do this, as the software processes the audio sequentially in 30-second chunks. It cannot look ahead to get the context of the full proceeding, and it cannot go back to fix things earlier in the transcript. This one-directional processing meant that if the AI misheard an unfamiliar term early in the proceeding, it was forced to make a guess and lock that error into the transcript. Even if the phrase was repeated perfectly clearly later in the transcript, the model had no ability to retroactively apply that newly learned context to correct its earlier mistake.

Overlapping speech

Another limitation we observed was how the model performed when two people speak at the same time. When a human-transcriber encounters overlapping speech, they can separate the voices, type out what each person said, and assign it to the correct speaker. Because Whisper processes audio as a single, flattened track and does not distinguish between different voices, it struggles to process two people speaking at once. We found that when this happened, Whisper would typically do one of two things: it would either completely omit the quieter speaker’s words and only transcribe the louder voice, or it would attempt to transcribe both speakers’ words together into a single, nonsensical run-on sentence. However, the latter was relatively rare.

Omitting quiet speakers

Whisper had a tendency to omit short passages from quiet speakers. This typically happened when a speaker was not near a microphone and only spoke briefly. The clearest example of this was the entry of a plea. When a magistrate asked the defendant how they pleaded, the defendant would often reply with a single word, ‘guilty’. Because the defendant was usually positioned away from the microphones, and the response was both quiet and very brief, Whisper frequently failed to transcribe it at all. Interestingly, the model had no trouble transcribing similarly quiet and unclear speech when that speech was longform. A defendant making a lengthy statement from the same position in the courtroom would generally be transcribed.

We assume this behaviour is likely due to bias in the model’s training data. Whisper was trained on 680,000 hours of audio collected from the internet, where the audio is presumably dominated by a single clear speaker (for example, podcasts, lectures, interviews, movies and video content). Short, quiet utterances in that training data would more often be background noise rather than meaningful speech that would appear in a transcript. The model appears to have learned to treat brief, quiet audio as something to ignore, even when it is in fact a substantive contribution to the proceeding. The practical implication for our research is that short responses from one party to another, for example a prosecutor or defence representative replying ‘that’s right’ to a question from the magistrate, can sometimes be missing from the transcript and need to be confirmed against the audio.

Errors in context

The 15 critical errors across the 30 cases must be contextualised against the size of the broader research project. The codebook for this CDVO research project captures more than 50 distinct variables for each case.¹¹

In this context, a critical transcription error typically affects only one or two of these variables. Across the 30 transcripts, 15 critical errors potentially distort a small fraction of the total fields that would be coded for those cases. The aggregate rate at which transcription error would cause a coding decision to change is very low and is likely within the margin of disagreement that is generally accepted for inter-rater reliability in social-science content analysis. While transcription error is not a substitute for, and should not be confused with, inter-coder error, it is reasonable to conclude that the practical impact of AI transcription error on the research dataset is small relative to the variation that already exists in any human-coded content analysis.¹²

Operational viability and cost

The third research question asked whether this approach can actually be operationalised within an agency setting. The evidence is that it can. The full pipeline runs on a standard issue laptop with 16 GB of RAM and no GPU, using freely available open-weights software, and processes audio at 2 to 3 times real-time. When comparing the cost of obtaining official transcripts, they can be up to 12 times more expensive, than obtaining audio files, which is much more cost effective, particularly where a large sample is required to ensure they are representative. This represents a substantial reduction in the cost of transcript-based research. The throughput is slow on CPU-only hardware, but the work can be run unattended in the background while researchers do other tasks, including overnight and during the weekend. A dedicated GPU would compress processing time dramatically but is not a prerequisite.

The information security considerations are equally important. Because Whisper is open-weights and runs entirely locally, no data is transmitted to a third party at any point. This is the feature that makes the approach viable for this type of data in the first place; a commercial cloud transcription service, however accurate, would not have satisfied relevant data management and privacy constraints for data classified as Protected.

AI-assisted workflow

The single clearest implication of the findings is that the AI-generated transcript should be treated as an aid to navigating the audio, rather than as a standalone record. The transcripts are accurate enough to support efficient research workflows: they let a researcher locate and skim the relevant parts of a 90-minute hearing in minutes rather than hours, and the great majority of substantive content is recorded faithfully. They are not, however, accurate enough to substitute for the audio itself in cases where the precise wording matters, and the risk of substantial omissions means that even a “clean-looking” transcript cannot be assumed to be complete. The workflow described in this section was designed specifically to capture the efficiency benefits of the transcript while preserving the audio as the source of truth.

To operationalise the AI transcripts, the research team built a custom Python script specifically for this project. The script synchronises the AI-generated text with the audio file using the model's output timestamps and presents a graphical interface that allows researchers to click on any line of text and immediately hear the corresponding audio playback.

The script works by parsing each line of the AI-generated transcript. As described earlier, each line of the raw output follows a consistent format that includes a start and end timestamp, a confidence score, and the transcribed text. For example:

¹¹ This included: the date the order was made, the date the order was breached, the conditions present on the order, the conditions that were breached, the type of aggrieved (aggrieved, named adult, or named child), whether the conduct constituted associated domestic violence, the type of conduct that constituted the breach (recorded across 28 individual variables grouped into five categories), and a wide range of contextual variables (physical harm, mental harm, financial harm, coercive control, weapon use, defendant mental health, defendant intoxication with alcohol or drugs, hardship of the sentence on the offender, vulnerability of the aggrieved, presence of a child, existence and relevance of the defendant's criminal and DV history, relationship between the aggrieved and the respondent, victim age, undeclared pre-sentence custody, pre-existing sentences or court orders, extra-curial penalties, and whether the defendant was in custody at sentence.

¹² Early research on error reduction in ASR (Automated Speech Recognition) models such as Whisper indicates that self-training mechanisms can reduce WER by 1-2%. See S N Dungavath, K S Nataraj and N Tiwari, 'Self-Training and Error Correction Using Large Language Models for Medical Speech Recognition' in 2024 *IEEE Conference on Engineering Informatics (ICEI)* (IEEE, 2024) 1. Self-training in machine learning is a semi-supervised learning approach where a model generates labels (or transcriptions, in the case of ASR) for previously unlabelled data, and those model-generated outputs (called 'pseudo-labels'), are then used as additional training data to further train the model.

[00:00:00.920 -> 00:00:03.120] [C:0.61] Mr Smith,
you are charged that on the 1st of January
[00:00:03.850 -> 00:00:06.920] [C:0.80] you
contravened the Domestic Violence Order.

When a researcher clicks on a line of text, the script processes the start timestamp and plays the corresponding section of the audio file from that point. The confidence score is used to colour each line of text on a gradient from red (low confidence) through to blue (high confidence). Anecdotally, researchers reported that segments with low confidence scores tended to contain a higher rate of transcription errors. The colour coding therefore acted as a visual flag, drawing the researcher's attention to the parts of the transcript most likely to need verification against the audio.

This AI-assisted workflow drastically improved research efficiency. Prior to the implementation of the AI-generated transcripts, researchers would have to scrub through the audio files. This is a tedious process of listening, guessing, and skipping to locate the relevant parts of the court proceeding. With the AI-generated transcripts, researchers were able to instead use the AI transcript as an interactive map to rapidly home in on the specific sections of the audio that deal with the substantive conduct of the offence.

In this workflow the AI-generated transcript augments the audio file rather than replacing it. This method mitigates the risk of AI hallucinations or critical semantic errors corrupting the research data, while preserving the time-saving benefits of text-based navigation.

Limitations of this paper

There are several limitations to bear in mind when interpreting the findings of this paper.

First, the parent sample of 100 cases contained no recordings classified as 'poor'. While the broader population of 1,395 cases is likely to contain such recordings, the AI's performance under those conditions was not measured. The error rates reported here should therefore be read as applicable to 'mostly clear' and 'partially clear' audio. Performance on poor-quality audio is expected to be substantially worse based on the gradient observed between the two quality categories that were represented, but the precise extent of that degradation is unknown.

Second, only 3 cases in the semantic sub-sample of 30 were classified as 'partially clear'. The audio-quality comparisons in the semantic analysis therefore rest on a small number of recordings and should be treated as indicative.

Third, no transcript was corrected by both researchers. Because each case was corrected by only one person,

formal inter-rater reliability could not be calculated, and observed differences between researchers could not be cleanly separated from differences in the cases assigned to each. Differences could reflect genuine differences in the accuracy of the transcripts assigned to each researcher but could also reflect differences in correction sensitivity.

Fourth, this evaluation tested a single model configuration (faster-whisper running large-v2, no initial prompt, no external VAD pre-processing). The performance of other Whisper variants, of other ASR systems, or of large-v2 with different configuration choices (for example, supplying an initial prompt with court-specific terminology) was outside the scope of this study.

Finally, the findings are drawn from one offence type (CDVO) in one jurisdiction (Queensland Magistrates Courts). Sentencing audio from other offence types or other courts may differ in vocabulary, formality, courtroom layout and recording infrastructure. The findings should not be assumed to generalise without further evaluation.

Conclusion

This paper set out to determine whether locally hosted AI transcription is a viable and cost-effective method for transcribing court proceedings for use in criminological research. The evidence collected supports a qualified yes.

On aggregate, Whisper transcribed 96.5% of words correctly across 100 Queensland sentencing hearings, with most cases falling well below a 5% word error rate. While 30% of the cases contained critical errors, the qualitative analysis confirmed that most of the errors observed would not often substantively affect the way a researcher would code a case as part of the broader study into the type of offending behaviour that occurred in sentenced charges of contravention of a domestic violence order.

Further, whilst 70% of cases in the semantic sub-sample contained no errors capable of misleading a researcher coding the substantive facts of the case, researchers were required to listen to the case audio in addition to using the transcripts to ensure accurate coding.

The dominant driver of error is audio quality rather than the time period of the recording, which means the model's performance on contemporary hearings is a

reasonable guide to its performance on hearings from previous time periods. The approach is operationally feasible. The full pipeline runs on a standard issue laptop, uses freely available open-weights software, and keeps all data inside the secure environment. Against a procurement cost for the audio files being up to 12 times less than for official court transcripts, when seeking a large scale, representative sample, the marginal cost of producing AI transcripts in-house is negligible.

The most important qualification is on the role the AI output should play. The transcripts are not accurate enough, and the risk of large sections being omitted are not detectable enough, for the AI transcript to be treated as a standalone record. They are, however, more than accurate enough to act as a navigation layer over the audio.

On that basis, AI transcription is recommended for adoption in future research projects of this type, where transcript-based navigation of audio recordings would materially reduce researcher time, and where verified coding decisions can be anchored against the audio itself.

Appendix 1: Codebook for semantic error analysis

If a cluster of incorrect words alters a single fact, concept, or continuous phrase (e.g., ‘clear of illicit drug uses’ turning into ‘clearly illicit drugs’), code it as one error event. If a cluster of incorrect words alter two distinct, unrelated facts, code them as multiple error events (e.g., ‘Miss [name] and the responder’ turning into ‘Mr [name] and the respondent’ changes two facts, the gender of one person, and the role of the other person).

For every error event, the coder must assign one severity rating (how much it impacts the transcript) and one theme (a classification for the error).

Part 1: Severity rating

Rating	Definition	Guidance
Minor	Non-substantive mistakes that do not alter the factual meaning of the sentence or disrupt readability.	An error occurred but the sentence is still readable and the meaning is the same. While an official court record would require this to be fixed, a general user might find it acceptable. Incorrect names are always classified as minor.
Moderate	Substantive mistakes that garble a sentence, hallucinate text, or misuse specific terminology. This includes errors that fundamentally alter, reverse, or obscure the facts, but do not mislead a researcher coding the data for this project.	This is the threshold where a general user might find the transcript unacceptable. The error erodes trust in the document and requires human interpretation to decipher the intended meaning. The error doesn't mislead a researcher trying to code the actual CDVO conduct for this project
Critical	Mistakes that fundamentally alter, reverse, or obscure the facts required for the research project.	The transcript is factually misleading. A researcher relying solely on the transcript would code the project variables differently.

Part 2: Themes

Theme	Definition	Examples
Grammar, syntax, lexical substitution, or correction	Shifts in sentence structure, punctuation, or swapping words that mean the exact same thing. Also includes speech correction, where the AI deviates from what was spoken verbatim to transcribe what was meant by correcting stutters, removing repeated text, or cleaning up speech dysfluency, such as a false start, repaired utterance, etc.	in a police car / in the police car in prison / imprisonment Is it updated history / Here is an updated history Think first before you show remorse / Think first before you and you showed remorse (AI cleaning up a stutter) she liked to help me / she liked help me (AI correcting grammar for a speaker for whom English was a second language)
Conversational filler	Errors involving courtroom pleasantries, human stuttering, or spoken conversational tags that carry no factual weight.	Omitting "Yes, your honour" ...officer? / ...I presume?
Proper nouns (Names/Places)	Misspelling, or mishearing specific names of people, towns, streets, or businesses.	Greensland / Greenslopes Marlborough / Maryborough
Domain-specific terminology	The AI struggles with domain-specific jargon or cultural terms, usually substituting them with phonetically similar words.	bury the bail / vary the bail the watch out / the watchhouse yarding circles / yarning circles

Theme	Definition	Examples
Phonetic garbling	The AI mishears the audio and substitutes a completely different word or phrase. It tries to make phonetic sense of the audio, resulting in bizarre or nonsensical text. Where garbling occurs as part of a term specific to criminal justice (e.g., aggrieved, DVO, SPER, etc) code instead to 'Domain-specific terminology'.	during the war / during the order prove it to yourself / privilege yourself You've done a fail / You are going to fail
Meaning reversal	The sentence is altered in a way that flips the narrative or reverses an outcome. Only applicable if it changes the facts of the matter, not for inconsequential reversals.	he had to walk away / you had to walk away clearly illicit drugs / clear of illicit drug uses Code this instead as 'phonetic garbling' as it doesn't reverse the substantive facts of the matter: We report to this man over here / You report to this man over here
Omission of content	Clauses, sentences, or paragraphs are dropped entirely. Only applies if information is lost, if a conversational filler (such as 'your honour') is omitted, code as 'conversational filler'. Often used where entire clauses, sentences or paragraphs are transcribed with the hallucinated text 'your honour'.	[omission] / It was just to say that he has been working that is all Thank you. Thank you. / There's no Queensland criminal history.
Hallucination	When the AI makes up text that was never spoken. Only used when the AI invents text when nothing is spoken. If the AI generates hallucinated text in place of actual spoken text, code as 'omission of content'.	Thank you. Thank you. / [nothing being spoken]

Queensland Sentencing Advisory Council

info@sentencingcouncil.qld.gov.au | sentencingcouncil.qld.gov.au | (07) 3738 9499

© Queensland Sentencing Advisory Council 2026.

This technical paper is licenced by Queensland Sentencing Advisory Council under a Creative Commons Attribution (CC BY) 4.0 International licence.

CC BY licence summary statement

In essence, you are free to copy, communicate and adapt this technical paper, as long as you attribute the work to the Queensland Sentencing Advisory Council.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>.

Content from this technical paper should be attributed as – 'Evaluating AI transcription of court proceedings for use in social science and legal research', Technical Paper No. 1, Queensland Sentencing Advisory Council, October 2026.

ISBN: 978-0-6452083-2-0



**Queensland Sentencing
Advisory Council**