Psychophysics of Duration Perception in Multimodal Large Language Models

Auss Abbood
Language Technology Lab
University of Cambridge
aa2613@cam.ac.uk

Zihao Fu
The Chinese University
of Hong Kong
zihaofu@cuhk.edu.hk

Nigel Collier
Language Technology Lab
University of Cambridge
nhc30@cam.ac.uk

Abstract

Studies of time perception in language models focus mainly on text-based settings. Yet multimodal models are increasingly used for tasks that require perceiving the passage of time, making the limits of their temporal perception important to understand. We measure the limits of duration perception from audio against a human baseline to address this gap. Drawing from psychophysics, we operationalise duration perception as a melody recognition task, testing when a melody becomes unrecognisable as it is slowed down. Of 21 tested models, most cannot recognise well-known melodies without labelled in-context examples. With in-context examples, seven recognise melodies reliably, most peaking within the median human recognition band. Generalisation is limited: (i) recognition peaks shift with in-context example speed, (ii) melody titles induce prediction bias, and (iii) providing examples with scientific pitch notation does not prevent accuracy declining at long note durations for most models. In humans, the recognition window is a property of the perceiver; in current models, it is largely a property of the prompt.

1 Introduction

There is increasing interest in understanding time perception in large language models (LLMs) (Chen et al., 2025). Existing approaches can be largely separated into two strands: One probes duration perception on the discourse level, measuring its alignment with humans' subjective judgements of duration or urgency (Wang et al., 2025; Cheng et al., 2026) ("*What feels longer, A or B?*"). The other strand investigates LLMs' ability to estimate task duration and calibration on the agentic level (Garikaparthi, 2026; Ofengenden and Andriushchenko, 2026) ("*How long do you need for this task?*"). However, the limits of time perception in multimodal large language models (MLLMs)—and how they compare with human limits—remain poorly understood. We investigate these limits using a melody recognition task that combines psychophysics and Gestalt psychology (Gilden, 2025).

Melody recognition survives temporal rescaling (Lerdahl and Jackendoff, 1996): a familiar tune remains recognisable at various pitches and tempi (Tenney and Polansky, 1980). Our perception limit is demarcated by when a familiar tune *stops* surviving temporal rescaling.

Figure 1 illustrates temporal grouping in melody recognition and contrasts human grouping limits with the unknown limits in models. At very short note durations, melody recognition fails due to physiological limitations (Hsieh and Saberi, 2007; Alves-Pinto et al., 2010; Uratani et al., 2014). However, above 1–2 s, melody recognition fails even though every note is perfectly audible (Warren et al., 1991; Andrews et al., 1998; Dowling et al., 2008); what dissolves is the tie between them (Fig. 1). This boundary defines our duration perception limits.

A transformer-based model has no obvious reason to have a perceptual ceiling similar to that of humans. It attends to the whole melody at once; recognising a slowed-down or sped-up melody is symbolically trivial. To this end, we assess 21 audio and omni models using the eight melodies of Warren et al. (1991) and verify our results on an expanded set of eleven more well-known melodies. None of the tested models reliably recognise the melodies from parametric knowledge alone at normal speed, so we could not use the same setup as the human study. We therefore provide labelled candidate melodies in context. Seven models qualify for the main experiment.

For most qualifying models, recognition initially peaks at note durations similar to those observed in humans. However, varying the speed of the in-context examples shifts this peak, with reliable

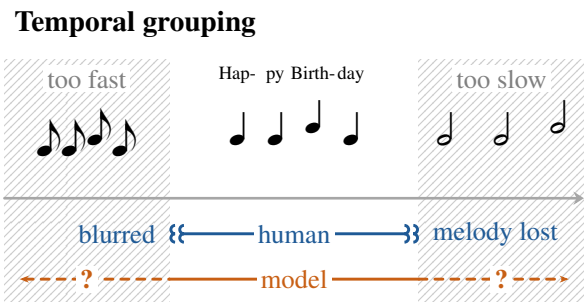


Figure 1: Conceptual illustration of temporal Gestalt grouping limits in humans and models. Blue and orange indicate human and model ranges, and hatched regions fall outside the human grouping range. Hooked endpoints denote approximate boundaries. Question marks indicate unknown model limits.

recognition confined to a narrow range around the examples' note durations. Melody labels also affect recognition. For example, replacing melody titles with neutral codes reduces prediction bias and improves accuracy. Together, these findings suggest that the apparent temporal perception limits depend strongly on the context provided, making it difficult to identify the models' intrinsic perceptual limits.

Our contributions can be summarised as follows:

1. We introduce a **psychophysical approach to measuring temporal perception limits in MLLMs**, using melody recognition to examine how duration affects recognition relative to a human baseline.
2. We show that **apparent temporal recognition limits shift with in-context example speed**, challenging their interpretation as stable perceptual boundaries.
3. We show that **melody labels can bias auditory recognition**, with neutral codes reducing prediction bias and improving accuracy.

2 Background

Time perception in LLMs Research on temporal reasoning in LLMs has overlooked time perception (Chen et al., 2025). Subsequent work filled this gap by operationalising time perception as understanding temporal relativity for example in context-dependent judgements tasks (Chen et al., 2025), or judging elapsed time in narrative settings (Bao and Srikumar, 2026). Limits in time perception are attributed to the need of mapping continuous time to discrete tokens (Wang et al., 2025).

Time perception limits keep resurfacing in agentic settings. Limits manifest as the inability to align tools calls with human time sensitivity (Cheng et al., 2026) or overestimating wall-clock time for agentic tasks by 4–7 times (Garikaparthi, 2026).

All studies find limitations in LLMs and agents to ground time perception through token alone. A possible explanation is that time perception requires multimodality or world models (Garikaparthi, 2026). Therefore, we focus on MLLMs.

Auditory psychophysics in LLMs Early psychophysical test on LLMs found merit in this method: they found that human perception properties can be recovered from LLMs trained on text alone (Marjieh et al., 2024).

As MLLMs become more established, perception tests are increasingly applied to multimodal data. In the auditory domain, SonicBench uses controlled stimuli to test absolute and comparative judgements of physical properties, including duration and tempo (Sun et al., 2026). Its stimuli are designed to be well above human discrimination thresholds, so it tests coarse judgements of individual acoustic properties rather than estimating model thresholds or limits of time perception.

Music-related benchmarks have overlap with psychophysical experiments in that they also test perception limits in MLLMs. Carone et al. (2026a), for example, test models on stimuli with varying rhythm, melody, and harmony. They find that models generalise poorly to transposed melodies presented as audio but perform near ceiling on MIDI (a symbolic music representation), concluding that models struggle to *listen* to music. The MUSE Benchmark test relational reasoning in music (Carone et al., 2026b). Among its tasks, it tests the recognition of grouping of beats into threes, fours, or fives. However, these are categorical judgements on a fixed stimulus set; MUSE does not vary note duration to estimate when temporal grouping breaks down.

Existing work therefore tests either physical duration judgements or musical grouping, but not the psychophysical limits of duration-dependent temporal grouping. Our work fills this gap.

Gestalt in LLMs Time perception in humans is commonly explained via internal-clock theories (van Rijn et al., 2014), such as Scalar Timing Theory, which posits the existence of an internal pacemaker whose pulses anchor the flow of time.

Gilden (2025) contrasts this theory with a Gestalt account in which temporal grouping explains phenomena not captured by internal-clock theory, such as the loss of temporal grouping shown in Fig. 1. Existing work on Gestalt and MLLMs has focused primarily on vision (Hou et al., 2026; Sha et al., 2025). Our study is the first to use a Gestalt account to test duration perception in MLLMs.

3 Duration perception probe

Our experimental setup is derived from the study by Warren et al. (1991). In short, it investigated the limits of temporal grouping in human temporal perception by letting humans classify well-known melodies at varying speeds. The point at which notes stop grouping into a melodic Gestalt and are instead perceived individually (Tenney and Polansky, 1980) denotes the perception limit. In humans, the median temporal perception window lies between 160 ms and 1280 ms (first quartile at 113 ms and third quartile at 2560 ms). The loss of temporal grouping at increased note durations, unlike the physiological limits at short ones, indicates a limit in our temporal perception.

The experiment is divided into three steps: training, screening and duration scaling. For each step, we describe our setup and compare it with the original human study.

Melodies We use the same eight melodies as the original study by Warren et al. (1991): Twinkle, Twinkle, Little Star, Happy Birthday, God Rest Ye Merry Gentlemen, Rock-a-bye Baby, Camptown Races, Yankee Doodle, Love Me Tender, and Skip to My Lou.

Each melody ℓ_i is represented as an ordered sequence of pitches,

$$\ell_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,n_i}), \quad n_i \in \{6, 7, 8, 9\}.$$

Melodies are presented isochronously, i.e., all notes in a stimulus have the same duration d_j, where $d_j \in D$. The set of possible note durations is

$$D = \{\mathrm{round}(40(\sqrt{2})^k)\,\mathrm{ms} : k \in \{0, \ldots, 12\}\} \cup \{3600\,\mathrm{ms}\}. \tag{1}$$

The final stimulus presented to the model is defined by its melody and duration,

$$s_{ij} = (\ell_i, d_j).$$

The melodies used, their scientific pitch notation, and an extended list of tested melodies are given in Appendix A.1. To avoid recognition by the first notes alone, the original study and this one use different entry points into a melody and then cyclically shift the sequence. For example, starting at the third pitch gives

$$(p_{i,3}, p_{i,4}, \ldots, p_{i,n_i}, p_{i,1}, p_{i,2}).$$

A melody may be looped, playing the same sequence of pitches n times. With a single loop, the sequence is played twice.

Melodies were rendered as sine waves. The exact setup is described in Appendix A.1.

Training This stage ensures that the tested subject can recognise the melody before any temporal dilation or contraction is applied. In the human study, melodies were played at a *normal* speed of 320 ms per note. Unless stated otherwise, normal speed refers to this value. Human participants needed to classify melodies correctly with the help of a list that contained the names of the tested melodies. Only if they recognised the melodies did they continue to the next stage.

We test two variations of the training phase for artificial intelligence (AI) models. The first variant presents melodies in a *zero-shot*, multiple-choice format, testing whether a model can recognise the melody from its parametric knowledge alone. The second variant frames the training phase as *in-context learning*, where we demonstrate the target and distractor melodies in context before asking the model to classify the target melody. This setting may bridge shortcomings in the model's ability to recognise a melody from its parametric knowledge alone.

Screening In the human study, participants qualified for the main experiment if they could classify melodies without the aid of a melody list. During screening, melodies were again played at normal speed, but now starting at the third or fourth note to rule out recognition by the first note alone. The melody was looped without pauses until the human subject made a classification. Subjects were included in the main study if they recognised six out of eight melodies.

We adopt this design, with an adaptation of the looping setup. An interactive classification as described in the human study is uncommon and difficult to implement in practice. Instead, we apply a static variation in which the target melody is looped once — the minimum used in the original study.

Limiting presentation to one loop also ensures comparability and caps test duration and costs. Reference melodies start at the first note and are played once. Each candidate set therefore contributes 16 trials (8 melodies $\times$ entry at note 3 or 4), and we scale the human acceptance criterion of 6/8 correct classifications to 12/16 trials.

Note that in the in-context version of the training phase, the training and screening phases are done in one go: providing in-context examples is the training phase and requesting models to make a prediction based on the looped melody starting from the third or fourth note is the screening phase.

Lastly, we test two formats for the screening experiment: *audio-only* and *text+audio*. The text+audio format uses a mix of text prompts with interleaved audio files for the reference and target melodies. Since not all models accept this input format, either because they accept audio input only, or because they are limited by the number of audio files they can process in a single prompt, we also translate all experiments into an audio-only format. Interleaved audio-text inputs can be challenging because audio produces long token sequences, and models must preserve the relationships between alternating textual labels and audio clips (Zhao et al., 2026). We use text-to-speech to translate the instructions to audio with minor adaptations of the instructions to fit the spoken setting. The audio-only setting produces a single audio file with spoken instructions and played melodies. Exact prompts and further details are in Appendix A.2.

Duration scaling In the human study, half the participants first heard the target melody at 40 ms per note. Note duration was then multiplied by $\sqrt{2}$ until the melody was recognised. The other half started at 3600 ms and note duration was divided by $\sqrt{2}$ until recognition.

During duration scaling, human participants heard each melody in a continuous loop until they made a classification. Model trials are independent, so we instead test every note duration in Eq. 1 and use a fixed presentation: each target melody starts at either the third or fourth note and is looped once. This gives two trials at each of the 14 note durations, or 28 trials per melody.

Musicians are able to solve this task with much higher accuracy, even at long note durations (Warren et al., 1991), because they can transcribe the melody as they listen to the recording, compensating for any perception limitations. We test whether LLMs are equally capable of overcoming such limitations with the help of transcription. In a transcription variant, we provide each candidate melody's scientific pitch notation (see Tabs. 2 & 3).

Evaluation Our experiments are designed as multiple-choice, asking the model to classify the target melody among the options given in context and to respond with that melody's name only. Because some melodies go by other names, we recognise alias names for melodies, too. For example, Twinkle, Twinkle, Little Star has the same melody as the ABC Song. We report the alias-corrected accuracy scores (more details in Appendix A.3). For the screening experiments, we test different numbers of in-context examples. To maintain comparability, we report chance-corrected accuracy in percent, defined as

$$100 \times \frac{\text{accuracy} - \text{chance}}{1 - \text{chance}},$$

where accuracy is the observed proportion correct and chance is the guessing probability determined by the number of in-context melodies. The chance-corrected 12/16 thresholds are 66.7%, 70.0%, and 71.4% for 4, 6, and 8 candidates, respectively.

We further analyse the detection envelope of LLMs to characterise how accuracy in melody-recognition changes with note duration. To this end, we model trial-level correctness using a Bernoulli generalised linear mixed model (GLMM) with a logit link. For stimulus s and model m, let $y_{s,m} \in \{0, 1\}$ indicate task success, and let d_s denote the note duration in seconds. We fit

$$\begin{aligned} y_{s,m} \mid p_{s,m} &\sim \text{Bernoulli}(p_{s,m}), \\ \text{logit}(p_{s,m}) &= \beta_0 + f(\log d_s) + u_m + v_{\ell[s]}. \end{aligned} \tag{2}$$

Here $f(\cdot)$ is a cubic B-spline basis expansion with K degrees of freedom, u_m is a random intercept for model identity, and $v_{\ell[s]}$ is a random intercept for melody identity. The random intercepts are assumed to follow

$$u_m \sim \mathcal{N}(0, \sigma_m^2), \qquad v_\ell \sim \mathcal{N}(0, \sigma_\ell^2).$$

We choose the spline degrees of freedom as described in Appendix A.4.

Models We test a variety of closed- and open-source models, some of which are multimodal or audio only. These include Gemini 3.1 Pro Preview (Google DeepMind, 2026), Gemini 3

Flash Preview (Google DeepMind, 2025b), Gemini 2.5 Pro and 2.5 Flash (Gemini Team, 2025), and Gemini 2.5 Flash-Lite (Google DeepMind, 2025a); GPT Audio (OpenAI, n.d.a) and GPT Audio Mini (OpenAI, n.d.b); Qwen3.5 Omni Plus, 3.5 Omni Flash (Qwen Team, 2026), 3 Omni Flash, 3 Omni 30B A3B Thinking and Instruct (Xu et al., 2025), 2 Audio 7B Instruct (Chu et al., 2024); Voxtral Small 24B 2507 (Liu et al., 2025); MiMo-V2 Omni and MiMo-V2.5 (Xiaomi, 2026b); Nemotron 3 Nano Omni 30B A3B Reasoning (NVIDIA, 2026); Audio Flamingo Next and Audio Flamingo Next Think (Ghosh et al., 2026a); Music Flamingo 2601 and Music Flamingo 2601 Think (Ghosh et al., 2026b). Throughout this paper, we use shortened model names. For example, Music Flamingo refers to Music Flamingo 2601.

Generation temperature is set to 0 where the API allows it. For Qwen3 Omni Thinking, we set the temperature to 0.6, Top P to 0.95, and Top K to 20 as recommended by the authors, since outputs at temperature 0 were very poor. MiMo-V2.5 does not support temperature control (Xiaomi, 2026a).

4 Results and Discussion

4.1 Screening

Our experiments aim to resolve three uncertainties.

(i) Is parametric knowledge alone enough to test models' ability to recognise melodies at varying speeds, or do we need to provide examples in-context? Our experiments show that models are barely able to recognise melodies at normal speed, starting from the first note, using parametric knowledge alone.

(ii) Is there a preferred input modality? Should instructions be audio only or is a mix of text and interleaved audio files more effective? We find that text+audio prompting yields much higher accuracy scores than audio-only.

(iii) How many in-context examples are adequate for the tested LLMs? We ablate the in-context variant, testing eight, six and four in-context examples.

Zero shot Results for both settings are in Tab. 4 in the Appendix. Most models in the text+audio setting classify only one out of eight melodies correctly, which is equal to closed-set chance. Qwen2 Audio and Audio Flamingo Think reach 25% accuracy, and Qwen3.5 Omni Plus reaches 38%.

In the audio-only setting, Gemini 3.1 Pro and Qwen3.5 Omni Flash reach 38% accuracy, and Gemini 3 and 2.5 Flash reach 25%. Despite music training, Music Flamingo and Audio Flamingo make at most one correct classification.

Overall, all models perform poorly and do not pass the screening phase, which requires models to classify $6/8$ melodies correctly.

In-context In this setting, models were presented with melodies in context before being asked to classify a melody (multiple-choice). For the four- and six-reference variants, we test seven candidate sets because audio models are sensitive to option selection and order (López et al., 2025).

Models perform much better in the text+audio setting than in the audio-only setting (Fig. 2). All models achieve higher overall scores and more models cross the selection threshold in some form. Both Gemini Pro models achieve accuracy close to 100%, even with varying numbers of references, unlike in the audio-only setting. Gemini 2.5 Flash also crosses the selection threshold with different numbers of candidate melodies. Qwen3.5 Omni Flash and Plus, Qwen3 Omni, and Gemini 3 Flash crossed the screening threshold in some of the experiments, including their whiskers. These models are identified by bold names (Fig. 2). Most models perform equally well for all three candidate numbers. Noticeable differences can be observed for Qwen3 Omni and Qwen3.5 Omni Plus. Note that Voxtral Small only has valid runs for the setting with four in-context examples because it cannot process more than five audio files.

In contrast, models in the audio-only screening experiment perform poorly, with only four models crossing the acceptance threshold in some form (Fig. 3). While both Gemini Pro models clear the criterion on average with four candidate melodies, both Gemini Flash models reach it only with their best candidate sets (whiskers) in the four-reference ablation. We can observe that model performance is mostly comparable among the different numbers of in-context examples. The exceptions are the two Gemini Pro models, which, on average, perform better with fewer in-context examples — even when comparing chance-corrected accuracy values. We find that Gemini Pro models deviate more often from the formatting instruction with eight reference melodies compared to six or four. We assume that this is driven by file length. A single experiment produces an almost three-minute audio file, which can dilute its opening instructions.

Therefore, the main experiments are conducted

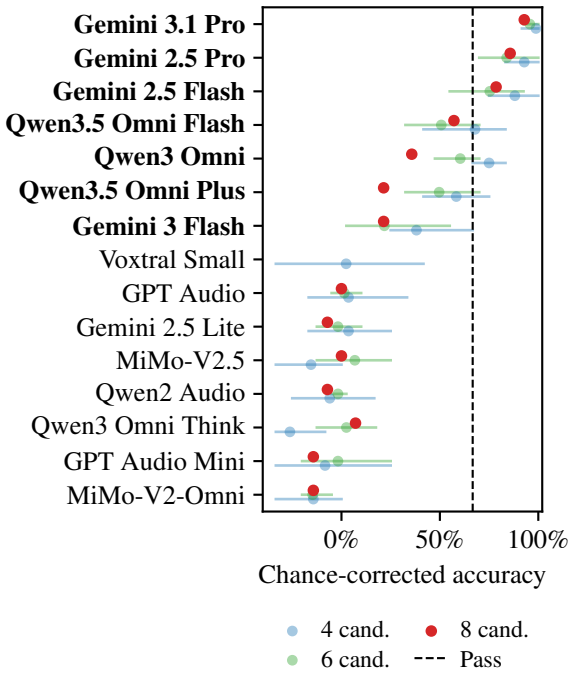


Figure 2: Chance-corrected text+audio screening accuracy with 4, 6, or 8 candidate melodies at 320 ms per note (Sec. 3). For 4 and 6 candidates, points and whiskers show the mean and min–max across seven balanced candidate sets. The dashed line marks the pass criterion, chance-corrected for 4 candidates. Bold model names identify models that meet the screening criterion in at least one condition, using the mean or best candidate set.

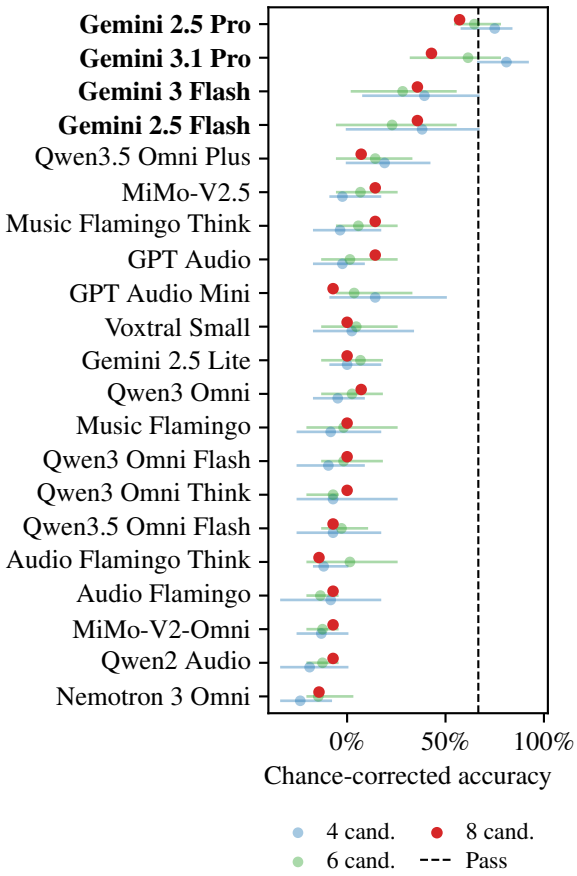


Figure 3: Chance-corrected audio-only screening accuracy under the design and encodings of Fig. 2. Bold model names identify models that meet the screening criterion in at least one condition, using the mean or best candidate set.

using in-context examples delivered as text+audio prompts. While the human study sought to identify a detection window, which requires strong melody recognition abilities, we are more interested in analysing how accuracy varies conditioned on the note duration of the target melody. We therefore pick all models that crossed the selection threshold, even if it was for a lower number of in-context examples or only the whiskers of the dot plots crossed the threshold (indicated by bold model names in Fig. 2). Still, to ensure comparability to the human study and given the overall similar results for different numbers of candidate melodies, we run the duration scaling experiments with eight in-context candidate melodies. Seven models pass screening and enter the duration scaling experiments: Gemini 3.1 Pro, Gemini 3 Flash, Gemini 2.5 Pro and Flash, Qwen3.5 Omni Plus and Flash, and Qwen3 Omni.

4.2 Duration scaling

Surprisingly, most models, though with reduced overall performance, peak around the human median perception window (Fig. 4). We summarise model performance over all trials and models using the GLMM described in Eq. 2. The broad posterior interval shown in the figure is driven by inter-model differences, most prominently by Gemini 3.1 Pro, which is the only model that maintains a high accuracy score regardless of note duration, though even this model shows a clear peak matching the median human recognition window.

All tested models except Gemini 3.1 Pro have difficulty confidently classifying a simple melody that varies in speed, though they perform much better when the in-context examples (played at 320 ms) and target melody are similar in speed. Performance for all models is poor if the target melody is played at very short note durations of less than 160 ms. A likely reason is that tokenisation is often performed in 160 ms chunks.

Gemini 3.1 Pro may be the only model with superhuman melody recognition. Even at very

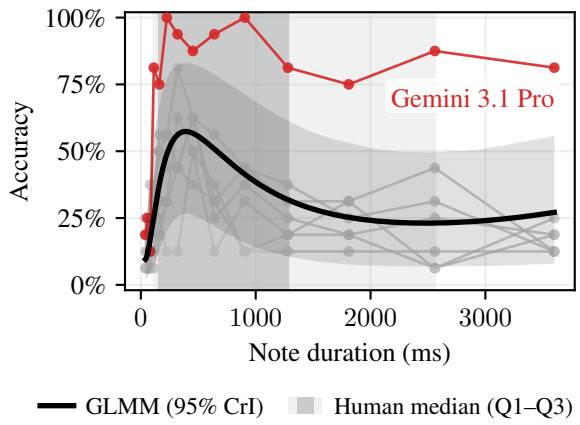


Figure 4: Accuracy versus target note duration for eight melodies (n=1568 trials). Targets begin at note 3 or 4 and are played twice. In-context references use 320 ms notes. Grey traces show the six other models that passed screening (Sec. 4.1). Red highlights Gemini 3.1 Pro. The black curve and band show the population-level GLMM estimate and 95% credible interval (CrI in the legend; Eq. 2). Grey bands mark median (dark) and quartile (light) human limits (Sec. 3).

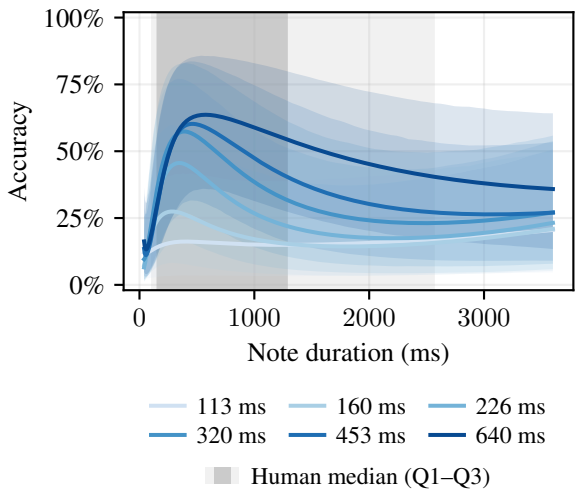


Figure 5: Population-level GLMM accuracy versus target note duration, one fit per in-context reference duration (113–640 ms; colours). Curves and bands show posterior means and 95% credible intervals. Human bands as in Fig. 4.

long note durations, where human perception shifts from a coherent melody to individual notes, it still achieves around 80% accuracy. However, the human interval is based on non-musicians and does not reflect the transcription-based strategy described in Sec. 3. Such a strategy could explain Gemini 3.1 Pro's plateau, but without its reasoning traces, we cannot distinguish transcription from superhuman temporal perception.

Ablation of in-context note duration Accuracy seems to peak near 320 ms (Fig. 4). To answer if in-context note duration influences peak melody recognition, we test in-context note durations going from 113–640 ms $\in D$, and refit the GLMM to assess how model performance behaves as a function of changed candidate melody speed. We can observe that overall model performance increases when we increase in-context melodies' note durations, and that the peak in accuracy shifts to the right (Fig. 5). On average, models are unable to make any correct prediction if in-context examples have a note duration of 113 ms, as for target melodies in the same speed in the main experiment.

Inter-melody differences Our work primarily compares model performance to the median human recognition band. This band differs across melodies. The median band for Happy Birthday (160–1810 ms) is wider than that for Skip to My Lou (226–905 ms) for example. Several factors may contribute, such as prior familiarity with a melody and melody-specific salience. We therefore also investigate inter-melody differences in model performance in Fig. 6. The figure shows Youden's J (Youden, 1950) and the false positive rate (FPR) for each model and melody for the human median recognition band (160–1280 ms) and above it (>1280 ms). Youden's J is defined as J = true positive rate − false positive rate. Assuming some melodies are easier and therefore over-predicted, Youden's J balances inflated true-positive rates against the FPR.

In Fig. 6, three phenomena emerge. (i) Twinkle, Twinkle, Little Star is over-predicted (high FPR). Gemini 2.5 Pro, Qwen3.5 Omni Plus and Flash, and Qwen3 Omni show increased FPR when the target melody is above the human median band. Qwen3 Omni already over-predicts it within the band. (ii) Some melodies are much harder to classify as note duration increases, but without an increase in FPR. This occurs for Rock-a-bye Baby and Camptown Races in Gemini 2.5 Pro, and Love Me Tender in Gemini 3.1 Pro and Gemini 2.5 Flash. (iii) Happy Birthday is well recognised with no significant FPR increase for Qwen3.5 Omni Plus.

A possible explanation for the over-prediction of Twinkle, Twinkle, Little Star in Fig. 6 is that it is an artefact of the in-context examples. We therefore vary their note duration and compare median FPR and median Youden's J for each of the six ref-

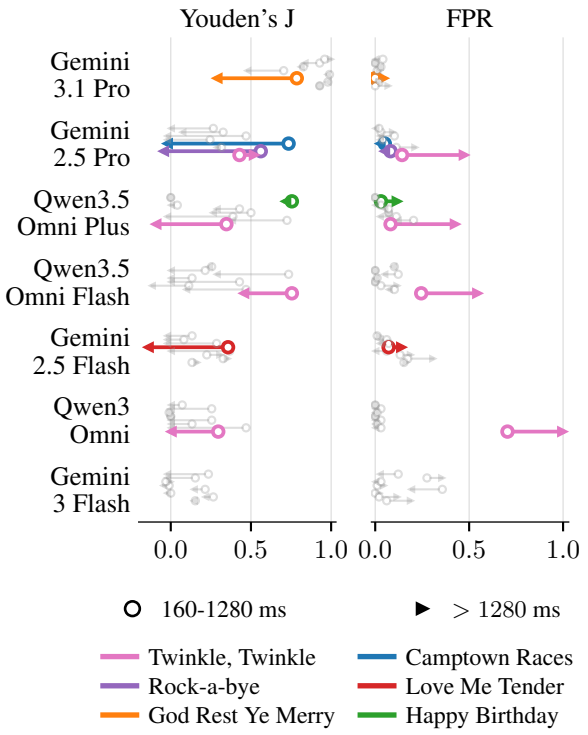


Figure 6: Per-model, per-melody changes in Youden's J and FPR from target note durations of 160–1280 ms (circles) to >1280 ms (arrowheads). Coloured arrows highlight atypical melody–model pairs, such as increasing FPR or declining Youden's J. Grey arrows show the remaining melody–model pairs.

erence durations (113–640 ms) at target durations above the human median band (Table 5). Twinkle's FPR is consistently higher, ranging from 0.36 to 0.60 compared to 0.03–0.05 for the other seven melodies. Its median Youden's J is likewise higher at every duration; the largest gap occurs at 320 ms (0.30 vs. 0.17).

There are various explanations why models would exhibit melody-specific biases as shown in Fig. 6. Commonly, such biases can be traced back to the training data. Since the models' training data are not accessible, we perform a surrogate data analysis on YouTube. Although the number of uploads matters more for training than view counts do, we also report view counts, since they may inform data selection.

We obtained video upload rates via the YouTube API across four fixed 24-hour publication windows in 2023–2024 in March and September. We exclude the winter months, assuming that God Rest Ye Merry Gentlemen could be uploaded more often around Christmas. Searches included alternative titles for each melody. Duplicate videos were

Melody	Videos	Channels	Median views
Twinkle Twinkle Little Star	122	103	4,105
Love Me Tender	47	42	95
Happy Birthday	36	34	56,925
Yankee Doodle	34	31	169
Rock-a-bye Baby	23	23	816
Skip to My Lou	15	15	105
God Rest Ye Merry Gentlemen	11	10	56
Camptown Races	6	6	116

Table 1: YouTube search results for eight melodies, pooled across four fixed 24-hour publication windows in 2023–2024: title-matched videos, unique channels, and median views per matched video. Aliases are pooled and duplicate videos removed.

removed within melodies, and title-based matching excluded irrelevant search results. We report the number of matching videos, unique uploading channels, and median views in Tab. 1.

Twinkle, Twinkle, Little Star has the most title matches (122) and unique uploading channels (103). With 4,105 views, it is the second-most-listened-to song in the analysis. The many uploading channels confirm genuine interest in the melody and argue against a few channels inflating upload counts. Happy Birthday ranks first in views (56,925) but has fewer title matches. Twinkle's high title-match count may reflect that the melody appears in songs under several aliases. If the model training data resemble YouTube and our query is representative, models likely encountered the melody more often than others, potentially explaining its higher FPR in Fig. 6. Because the melody has different names, deduplication may not have removed many training instances. God Rest Ye Merry Gentlemen and Camptown Races also became harder to classify as note duration increased; our YouTube query finds the fewest videos for these melodies (eleven and six, respectively).

We explore whether the melody bias can be mitigated by replacing music titles with neutral letter–digit codes, such as "D4". Indeed, over-prediction of melodies decreases and overall performance increases. Fig. 7 shows that no melody has an overly high recall or precision. In comparison, prompting models with the full melody name leads to an over-prediction of Twinkle, Twinkle, Little Star for some models. The model performance is also

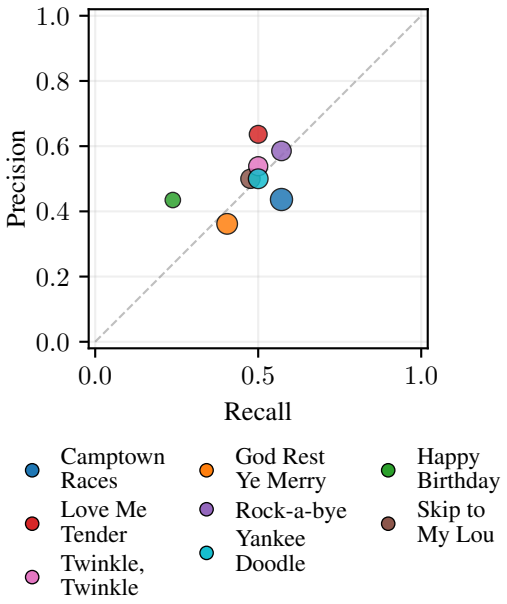


Figure 7: Per-melody precision and recall in the neutral-code ablation, pooled over the trials of all screened models at target note durations >1280 ms. Marker area scales with how often the title was predicted. The dashed diagonal denotes equal precision and recall.

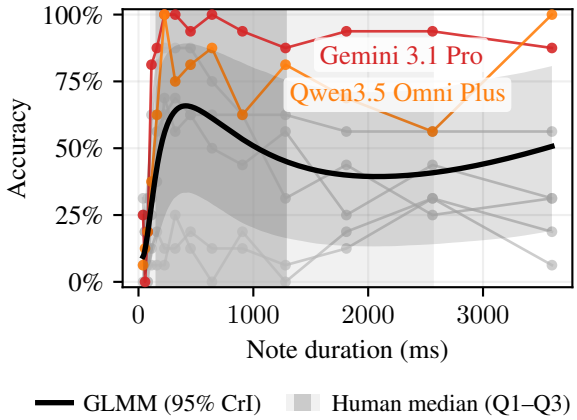


Figure 8: Accuracy versus target note duration when melody titles are replaced with neutral letter–digit codes. The setup, model selection, and encodings follow Fig. 4. Red and orange highlight Gemini 3.1 Pro and Qwen3.5 Omni Plus, respectively.

better when using codes instead of melody titles.

The GLMM fitted curve peaks above 60% (Fig. 8), versus around 50% with full titles (Fig. 4). Gemini 3.1 Pro and Qwen3.5 Omni Plus lie above the fit's credible interval.

More melodies While eight melodies may yield robust results about human time perception, inter-melody bias findings as observed in Fig. 6 cast doubt on the robustness of our results. Therefore, we add eleven more songs that are also considered to be widely known (in the English-speaking world), ensuring that melody recognition is only constrained by note duration, not by familiarity with the melody. These include America, Frère Jacques, Here Comes the Bride, Jingle Bells, Jolly Good Fellow, London Bridge, Mary Had a Little Lamb, Old MacDonald, Row, Row, Row Your Boat, The Star-Spangled Banner, and Three Blind Mice (Gfeller et al., 2002; Kang et al., 2009).

Extending the set beyond the eight melodies of Warren et al. (1991) reproduces the same pattern (Appendix, Fig. 15): all models peak in the same region with similar accuracy, while Gemini 3.1 Pro shows a similar but more stable peak and plateau.

Scientific pitch notation We next report the transcription ablation introduced in Sec. 3. When we supply scientific pitch notation for each melody, average peak accuracy rises from just above 50% to around 60% (Figs. 4 and 9). Gemini 3.1 Pro remains near-perfect even at 3600 ms per note. The other models improve but still decline with duration. Thus, notation helps most models, but only Gemini 3.1 Pro sustains high long-duration accuracy, as human musicians do; whether it transcribes internally remains unclear.

5 Conclusion

In this work, we test the limits of duration perception in MLLMs using a melody-recognition task. Most of the 21 models tested struggle to recognise familiar melodies, even with labelled in-context examples. The few that succeed peak within the human recognition band. However, this resemblance does not necessarily reflect a comparable perceptual limit, as the recognition window shifts with the speed of the references. Recognition also depends on what a melody is called: replacing familiar titles with neutral codes reduces prediction bias and improves performance. Even with scientific pitch notation, most models cannot sustain recognition at long note durations as human musicians do. Gemini 3.1 Pro is an exception, recognising melodies even at 3600 ms per note, though whether this reflects greater perceptual capacity or reasoning strategies remains unclear. MLLMs have the potential to surpass human perceptual limits, and this work begins to stake out these boundaries.

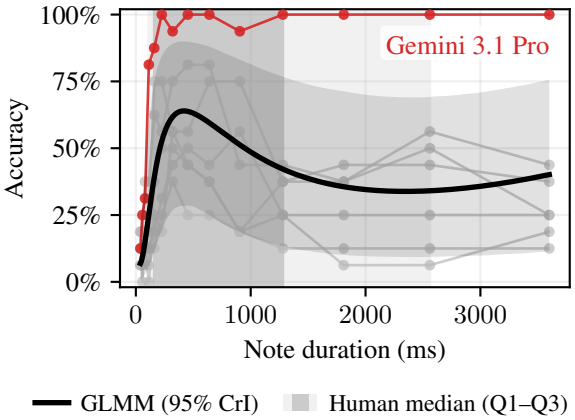


Figure 9: Accuracy versus target note duration with scientific-pitch notation for each in-context reference. The setup, model selection, and encodings follow Fig. 4.

6 Limitations

Human participants heard melodies in interactive loops and relied on prior familiarity, whereas models received fixed trials and in-context references because their zero-shot recognition was insufficient. The resulting curves should therefore be interpreted as behavioural patterns rather than as directly equivalent perceptual thresholds. Once interactive models become the norm, we should test them more similarly to the human setup. Furthermore, our stimuli are isochronous sine-tone versions of familiar melodies drawn largely from English-speaking musical culture, so the findings may not generalise to natural recordings, unfamiliar or non-Western music, or duration perception beyond melodic grouping. Finally, black-box behaviour cannot reveal the mechanisms underlying model performance. Sustained accuracy may arise from temporal grouping, pitch tracking, transcription, or retrieval from training data; the lack of access to these data prevents us from distinguishing among these explanations.

Acknowledgements

Auss Abbood was supported by the Economic and Social Research Council [ES/Y001788/1]. This work was supported by the Google Cloud Research Credits program.

References

Ana Alves-Pinto, Sylvie Baudoux, Alan R. Palmer, and Christian J. Sumner. 2010. Forward masking estimated by signal detection theory analysis of neuronal responses in primary auditory cortex. *JARO: Journal of the Association for Research in Otolaryngology*, 11(3):477–494.

Melinda W. Andrews, W. Jay Dowling, James C. Bartlett, and Andrea R. Halpern. 1998. Identification of speeded and slowed familiar melodies by younger, middle-aged, and older musicians and nonmusicians. *Psychology and Aging*, 13(3):462–471.

Catherine Bao and Vivek Srikumar. 2026. The machine's internal clock: Do LLMs share human temporal illusions? ArXiv:2608.15394 [cs.CL].

Brandon James Carone, Iran R. Roman, and Pablo Ripollés. 2026a. LLMs can read music, but struggle to hear it. an evaluation of core music perception tasks. In *Proceedings of the 1st International Workshop on Emerging AI Technologies for Music*, volume 303 of *Proceedings of Machine Learning Research*, pages 1–27. PMLR.

Brandon James Carone, Iran R. Roman, and Pablo Ripollés. 2026b. The MUSE benchmark: Probing music perception and auditory relational reasoning in audio LLMs. In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16097–16101, Barcelona, Spain.

Shuang Chen, Yining Zheng, Shimin Li, Qinyuan Cheng, and Xipeng Qiu. 2025. Perceive the passage of time: A systematic evaluation of large language model in temporal relativity. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8304–8313, Abu Dhabi, UAE. Association for Computational Linguistics.

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, and Soheil Feizi. 2026. Your LLM agents are temporally blind: The misalignment between tool use decisions and human time perception. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37082–37104, San Diego, California, United States. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and

Jingren Zhou. 2024. Qwen2-Audio technical report. ArXiv:2407.10759 [eess.AS].

W. J. Dowling, J. C. Bartlett, A. R. Halpern, and M. W. Andrews. 2008. Melody recognition at fast and slow tempos: Effects of age, experience, and familiarity. *Perception & Psychophysics*, 70(3):496–502.

Aniketh Garikaparthi. 2026. Can LLMs perceive time? an empirical investigation. ArXiv:2604.00010 [cs.CL]. Presented at the I Can't Believe It's Not Better Workshop at ICLR 2026.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Google technical report. arXiv:2507.06261. Accessed: 2026-09-13.

Kate Gfeller, Christopher Turner, Maureen Mehr, George Woodworth, Robert Fearn, John F Knutson, Shelley Witt, and Julie Stordahl. 2002. Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults. *Cochlear Implants International*, 3(1):29–53.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026a. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. ArXiv:2604.10905 [cs.SD].

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang-gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2026b. Music flamingo: Scaling music understanding in audio language models. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

David Gilden. 2025. *The Psychology of Human Temporality*. Cambridge University Press, Cambridge.

Google DeepMind. 2025a. Gemini 2.5 Flash-Lite Model Card. Accessed: 2026-09-13.

Google DeepMind. 2025b. Gemini 3 Flash Model Card. Accessed: 2026-09-13.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. Accessed: 2026-09-13.

Wenjin Hou, Wei Liu, Han Hu, Xiaoxiao Sun, Serena Yeung-Levy, and Hehe Fan. 2026. Seeing is believing? a benchmark for multimodal large language models on visual illusions and anomalies. ArXiv:2602.01816 [cs.CV].

I-Hui Hsieh and Kourosh Saberi. 2007. Temporal integration in absolute identification of musical pitch. *Hearing Research*, 233(1-2):108–116.

Robert Kang, Grace Liu Nimmons, Ward Drennan, Jeff Longnion, Chad Ruffin, Kaibao Nie, Jong Ho Won, Tina Worman, Bevan Yueh, and Jay Rubinstein. 2009. Development and validation of the university of washington clinical assessment of music perception test. *Ear & Hearing*, 30(4):411–418.

Fred Lerdahl and Ray S. Jackendoff. 1996. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, MA. Reissue, with a new preface. Originally published 1983.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde

Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. 2025. Voxtral. ArXiv:2507.13264 [eess.AS].

Fernando López, Santosh Kesiraju, and Jordi Luque. 2025. Robustness assessment of large audio language models in multiple-choice evaluation. ArXiv:2510.04584 [cs.CL]. Accepted at Interspeech 2026.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 2024. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445.

NVIDIA. 2026. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. ArXiv:2604.24954 [cs.LG].

Michael Ofengenden and Maksym Andriushchenko. 2026. Your agents are not time aware. LessWrong.

OpenAI. n.d.a. GPT-Audio. OpenAI API model documentation. Accessed: 2026-09-13.

OpenAI. n.d.b. GPT-Audio Mini. OpenAI API model documentation. Accessed: 2026-09-13.

Qwen Team. 2026. Qwen3.5-Omni technical report. ArXiv:2604.15804 [cs.CL].

Hedderik van Rijn, Bon-Mi Gu, and Warren H. Meck. 2014. Dedicated clock/timing-circuit theories of time perception and timed performance. In Hugo Merchant and Victor de Lafuente, editors, *Neurobiology of Interval Timing*, volume 829 of *Advances in Experimental Medicine and Biology*, pages 75–99. Springer New York, New York, NY.

Jingyuan Sha, Hikaru Shindo, Kristian Kersting, and Devendra Singh Dhami. 2025. Gestalt vision: A dataset for evaluating gestalt principles in visual perception. In *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning (NeSy)*, volume 284 of *Proceedings of Machine Learning Research*, pages 873–890. PMLR.

Yirong Sun, Yanjun Chen, Xin Qiu, Gang Zhang, Hongyu Chen, Daokuan Wu, Chengming Li, Min Yang, Dawei Zhu, Wei Zhang, and Xiaoyu Shen. 2026. SonicBench: Dissecting the physical perception bottleneck in large audio language models. ArXiv:2601.11039 [cs.SD].

Ole Tange. 2026. GNU Parallel 20260422 ('Artemis II').

James Tenney and Larry Polansky. 1980. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241.

Yuka Uratani, Tadashi Nishimura, Seiji Nakagawa, Tadao Okayasu, Toshiaki Yamanaka, and Hiroshi Hosoi. 2014. Suppression of subsequent N1m amplitude when the masker frequency is different from the signal. *Journal of Experimental Neuroscience*, 8:1–5.

Minghan Wang, Ye Bai, Thuy-Trang Vu, Ehsan Shareghi, and Gholamreza Haffari. 2025. Discrete minds in a continuous world: Do language models know time passes? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18703–18729, Suzhou, China. Association for Computational Linguistics.

Richard M. Warren, Daniel A. Gardner, Bradley S. Brubaker, and James A. Bashford, Jr. 1991. Melodic and nonmelodic sequences of tones: Effects of duration on perception. *Music Perception*, 8(3):277–289.

Xiaomi. 2026a. Model hyperparameters. Xiaomi MiMo API Open Platform documentation.

Xiaomi. 2026b. Xiaomi MiMo: Model Release. Model release log. Entries for MiMo-V2-Omni (March 18, 2026) and MiMo-V2.5 (April 23, 2026). Accessed: 2026-09-13.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian

Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025. Qwen3-Omni technical report. ArXiv:2509.17765 [cs.CL].

W. J. Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Tong Zhao, Chenghao Zhang, Yutao Zhu, and Zhicheng Dou. 2026. ATIR: Towards audio-text interleaved contextual retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22032–22046, San Diego, California, United States. Association for Computational Linguistics.

Melody	Notes
Twinkle Twinkle Little Star	F#4 F#4 C#5 C#5 D#5 D#5 C#5 C#5
Happy Birthday	G#4 G#4 A#4 G#4 C#5 C5
God Rest Ye Merry Gentlemen	G4 G4 D5 D5 C5 Bb4 A4 G4
Rock-a-bye Baby	E4 G4 E5 D5 C5 E4 G4 C5 B4
Camptown Races	C#5 C#5 A#4 C#5 D#5 C#5 A#4
Yankee Doodle	Bb4 Bb4 C5 D5 Bb4 D5 C5
Love Me Tender	F4 Bb4 A4 Bb4 C5 G4 C5
Skip to My Lou	B4 B4 G4 G4 B4 B4 D5 D5

Table 2: The original eight melodies and their scientific-pitch sequences, from Warren et al. (1991).

Melody	Notes
America	G4 G4 A4 F#4 G4 A4
Frere Jacques	F4 G4 A4 F4 F4 G4 A4 F4
Here Comes the Bride	F4 Bb4 Bb4 Bb4 F4 C5 A4 Bb4
Jingle Bells	A4 A4 A4 C5 F4 G4 A4
Jolly Good Fellow	A4 A4 A4 G4 A4 Bb4 A4
London Bridge	A4 B4 A4 G4 F#4 G4 A4
Mary Had a Little Lamb	B4 A4 G4 A4 B4 B4 B4
Old Mac Donald	A4 A4 A4 E4 F#4 F#4 E4
Row Row Row	D4 D4 D4 E4 F#4 F#4 E4 F#4
Star-Spangled Banner	A4 F#4 D4 F#4 A4 D5
Three Blind Mice	F#4 E4 D4 F#4 E4 D4

Table 3: Eleven widely familiar melodies added to the original set, with their scientific-pitch sequences. Selection is based on Gfeller et al. (2002) and Kang et al. (2009).

A Experiment setup

We used GNU Parallel (Tange, 2026) to run experiment sweeps across models and experimental conditions.

A.1 Melodies

Notes in melodies were rendered as sine waves at a sampling rate of 44.1 kHz, with a peak amplitude of 0.2, 5 ms linear attack and release ramps to avoid onset/offset clicks, and a 20 ms silent gap after each sounding portion. Note durations are specified onset-to-onset (i.e., each interval includes both the tone and the following gap). The resulting waveforms were saved as mono WAV files.

A.2 Prompts

This section contains the prompts used for the screening phase, which include audio-only and text+audio formatted prompts in a zero-shot setting (Figs. 10 & 11) and an in-context setting (Figs. 12 & 13). NVIDIA's Audio Flamingo models were prompted differently in the audio-only setting. According to their model card, the model should not be prompted with only an audio file. Therefore, we prompt them with the instruction `You will receive one audio file for a melody recognition task. The file contains spoken instructions, optional labelled reference melodies, and then a target melody. Listen to the full audio. Classify only the final target melody. Answer with exactly one title from the reference titles mentioned.` before supplying the WAV file.

Spoken instructions use OpenAI's `gpt-4o-mini-tts` (voice `alloy`). They are concatenated with the labelled reference clips and the target melody, with 500 ms of silence between consecutive segments.

Zero-shot prompt (text+audio)

```
You are participating in a melody
recognition task. The possible answer
titles are: Skip to My Lou. Happy
Birthday. [...]. God Rest Ye Merry
Gentlemen. You will hear one melody.
Answer with only the title of the
melody, no extra words.
<: target audio />
```

Figure 10: Zero-shot text+audio prompt used for screening. Text instructions precede the target audio. Candidate-title order is randomised per trial.

Zero-shot prompt (audio-only)

```
<: spoken instructions (TTS)>
  You are participating in a melody
  recognition task. [...] The melody
  starts now.
</: spoken instructions (TTS)>
<: target audio />
```

Figure 11: Zero-shot audio-only prompt used for screening. Text-to-speech instructions and target audio are concatenated into one WAV. Differences from Fig. 10 are bold. Candidate-title order is randomised per trial.

Few-shot prompt (text+audio)

You are participating in a melody-recognition screening task. The labelled examples in this prompt are the only possible answers for this trial. The target melody is guaranteed to be one of them.
Each labelled example is wrapped between two text markers: "Reference: <title>" immediately before its audio and "End Reference: <title>" immediately after. The audio you must classify comes last and is preceded by "Classify this audio — choose the best matching title above:". The target may start at a later note and may loop, so do not rely on exact file equality. Choose the title of the reference that best matches the target.
Return only the melody title, exactly as written below, with no extra words.
Eligible answers:
Happy Birthday
Camptown Races
[...]
Skip to My Lou
Reference: Twinkle Twinkle Little Star
<🔊: reference audio>
End Reference: Twinkle Twinkle Little Star
Reference: Happy Birthday
<🔊: reference audio>
End Reference: Happy Birthday
[...]
Classify this audio — choose the best matching title above:
<🔊: target audio />

Figure 12: In-context text+audio prompt used for screening and duration-grid experiments. The displayed reference list is abbreviated with [...]. Experiments use eight references unless noted. The prompt remains fixed while the target clip's melody, entry point, and/or note duration varies.

The main grid experiments reuse the in-context text+audio prompt (Fig. 12) with different note durations for the target melody. The ablation that replaces melody titles with neutral letter–digit codes is identical to Fig. 12 except for the melody titles. The ablation prompt that provides the scientific pitch notation (Fig. 14) extends the main prompt (Fig. 12). Differences are highlighted in the Figure.

Few-shot prompt (audio-only)

<🔊: spoken instructions (TTS)>
You are participating in a melody recognition task.
[...]
The possible answer titles are: Happy Birthday. Camptown Races. [...]. Skip to My Lou.
Reference example 1. The title is Twinkle Twinkle Little Star**. The reference melody starts now.**
</🔊: spoken instructions (TTS)>
<🔊: reference audio>
<🔊: spoken instructions (TTS)>
That was reference example 1, titled Twinkle Twinkle Little Star**.**
[...]
Now classify the target melody. The target melody starts now.
</🔊: spoken instructions (TTS)>
<🔊: target audio />

Figure 13: In-context audio-only prompt used for screening. Text-to-speech instructions, labelled reference clips, and target audio are concatenated into one WAV. Differences from Fig. 12 are bold. Experiments use eight references unless noted.

A.3 Scoring

A trial is correct only if the response names exactly one melody from the closed set for that trial. Matching is case-insensitive, ignores punctuation, and uses word boundaries, so a title must appear as a whole phrase. We also accept pre-specified aliases: Twinkle, Twinkle, Little Star (ABC Song, Alphabet Song, Baa Baa Black Sheep), Happy Birthday (Happy Birthday To You), God Rest Ye Merry Gentlemen (God Rest You Merry Gentlemen, with or without a comma), and Love Me Tender (Aura Lee, Aura Lea). All other tunes are scored only on their canonical title. For thinking models, we discard <think>...</think> and score only the remainder. An unclosed think block, no recognised title, or two or more distinct titles count as incorrect. In the neutral-code ablation, the same rule applies to the letter-digit codes, not the real names.

Few-shot prompt (text+audio, note sequences)

You are participating in a melody-recognition screening task.
[...]
Each reference audio is followed by its note sequence in scientific pitch notation (e.g. C#5), which describes that reference tune.
[...]
Return only the melody title, exactly as written below, with no extra words.
Reference: Twinkle Twinkle Little Star
<🎵: reference audio>
This melody's notation is: F#4 F#4 C#5 C#5 [...]
End Reference: Twinkle Twinkle Little Star
Reference: Happy Birthday
<🎵: reference audio>
This melody's notation is: G#4 G#4 A#4 G#4 [...]
End Reference: Happy Birthday
[...]
Classify this audio — choose the best matching title above:
<🎵: target audio />

Figure 14: In-context text+audio prompt with scientific-pitch sequences appended to each reference. Text shared with Fig. 12 is elided; additions are bold.

A.4 GLMM

We fit the duration GLMM in Eq. (2) with four NUTS chains (1000 warmup iterations and 1000 draws, `target_accept`=0.95, seed 0) and compare cubic B-splines with $K \in \{3, 4, 5, 6, 7\}$ by PSIS-LOO on the main eight-melody grid (n=1568 trials). K=5 has the highest ELPD; K=4 is within one standard error of that fit (ΔELPD=2.7, SE of the difference =3.2) and is the smallest such K, so all reported curves use K=4. The shaded band is the 95% posterior interval of population-level accuracy after integrating out the model and melody intercepts.

B Further results

B.1 Extended melody list

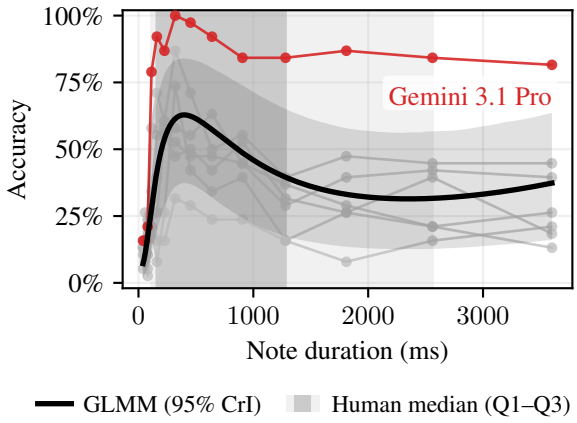


Figure 15: Accuracy versus target note duration for the extended set of 19 melodies. The setup, model selection, and encodings follow Fig. 4.

B.2 Zero-shot screening

	Acc. (%)	
Model	Text+audio	Audio-only
Qwen3.5 Omni Plus	38%	25%
Gemini 3.1 Pro	12%	38%
Qwen3.5 Omni Flash	12%	38%
Gemini 2.5 Flash	12%	25%
Gemini 3 Flash	12%	25%
Audio Flamingo Think	25%	0%
Gemini 2.5 Pro	12%	12%
GPT Audio Mini	12%	12%
Qwen2 Audio	25%	0%
Qwen3 Omni	12%	12%
Qwen3 Omni Think	12%	12%
Audio Flamingo	0%	12%
GPT Audio	12%	0%
MiMo-V2-Omni	12%	0%
MiMo-V2.5	12%	0%
Music Flamingo	12%	0%
Music Flamingo Think	12%	0%
Nemotron 3 Omni	0%	12%
Qwen3 Omni Flash	0%	12%
Gemini 2.5 Lite	0%	0%
Voxtral Small	0%	0%

Table 4: Zero-shot screening accuracy over eight melody trials per modality (eight-way closed set; chance =12.5%). Targets begin at the first note. **Text+audio** uses text instructions followed by target audio. **Audio-only** concatenates text-to-speech instructions and target audio into one WAV.

B.3 Over-prediction of Twinkle, Twinkle, Little Star

	FPR		Youden's J	
Ref. (ms)	Twinkle	Others	Twinkle	Others
113	0.60	0.03	0.17	0.07
160	0.53	0.04	0.18	0.06
226	0.48	0.05	0.24	0.15
320	0.41	0.04	0.30	0.17
453	0.41	0.05	0.21	0.19
640	0.36	0.04	0.33	0.27

Table 5: Median FPR and median Youden's J at target note durations >1280 ms, pooled over the trials of all screened models, comparing Twinkle, Twinkle, Little Star with the median of the remaining seven melodies, by in-context reference note duration. Higher FPR indicates stronger over-prediction; higher Youden's J indicates better discrimination.