AIMC 2023

The Ai Music Generation Challenge 2022: Summary and Results

Bob L. T. Sturm

Published on: Aug 29, 2023

URL: https://aimc2023.pubpub.org/pub/ynxz9uu7

ABSTRACT

We discuss the design and results of *The Ai Music Generation Challenge 2022* and compare it to the previous two challenges. While the 2020 challenge focused on generating Irish double jigs, and the 2021 challenge focused on generating Swedish slängpolskor, the 2022 challenge posed three sub-challenges in the context of Irish traditional music: generation of *reels*, judging tune submissions, and titling tunes. In total seven systems participated in the sub-challenges, along with benchmark systems. One tune was awarded first prize by the judges, and two tunes shared second prize. A submitted system for judging tunes clearly performed better than two benchmarks. Finally, human tune-titling outperformed the benchmark and submitted system, but gave rise to some interesting issues about tune titling.

Introduction

The Ai Music Generation Challenge – a public machine learning challenge aiming to motivate research in the application of machine learning research to living traditional music practices – has had three iterations now. The first challenge in 2020 [1] focused on the Irish double jig. The second challenge the following year [2] focused on the Swedish slängpolska. The 2022 challenge [3] focused on another form of Irish traditional dance music: the *reel*, a multipart tune performed with duple meter. Each part has 8 bars and is typically repeated twice. Figure 1 shows one example of a reel, which is a tune that is still played today but with different ornamentation and phrasing, not to mention many variations.

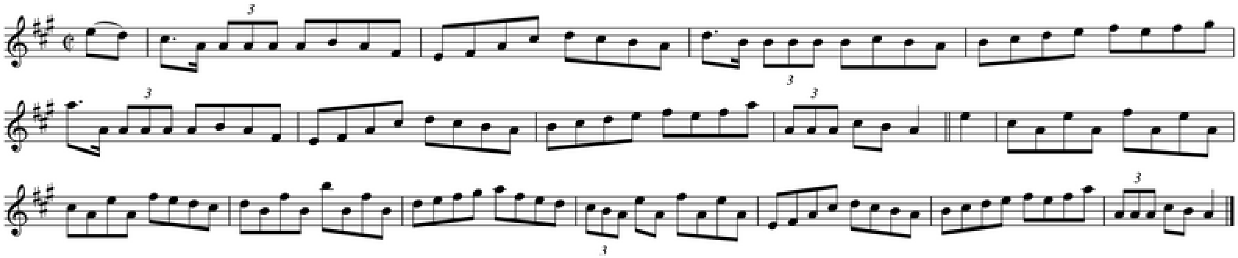


Figure 1: ``The Mason's Apron'' (no. 598), as printed in O'Neill's `1001’ (1907). Hear here Irish fiddler Kevin Burke and accordionist Jackie Daly perform a version of this tune in 1978.

A major change in *The Ai Music Generation Challenge 2022* from the former ones is its inclusion of three “sub-challenges”. The first sub-challenge is to generate reels in the style of the 350 in O’Neill’s 1001 [4]. The second sub-challenge is to generate scores for submitted reels, which are compared against the scores of the judges. The third sub-challenge is to generate titles for submitted reels, which are assessed by the experts for how well they fit the tunes. Participants could enter one system into each of the three challenges. In the following, we review the design and results of the 2022 challenge. Several submitted tunes are notated with reflections of the judges, showing a diversity of approaches and complex relationships with extra-musical factors. We discuss several points of comparison with the previous challenges, and look forward to the challenge of the coming year.

This paper contributes a thorough accounting of the unique event, presenting a picture of a contemporary interfacing of traditional music practice and AI music research. Even though this paper is mostly descriptive, I hope that in 60 years it provides a fair and honest picture of this event and its participants – which is unique in how it is engaging traditional music practitioners with the outcomes of AI music research, and vice versa.

The Ai Music Generation Challenge 2022

Design of three sub-challenges

The design of the 2022 *generation* sub-challenge follows that of the 2021 challenge. Each participant is required to submit a collection of 1,000 tunes generated by their system, along with a brief document describing their system, and the selection of a generated tune that they want evaluated by the judges. The tunes in a submission must be rendered as MIDI and notation (such as abc, musicXML, or staff). From a submitted collection, the organizers (we) select nine tunes at random and combine with the elected tune. We then render the ten tunes as staff notation (PDF), MIDI file, and as an MP3 (using the same appropriate tempo and a piano soundfont). The four human judges are sent the entire collection at once for evaluation.[1]

The judging of the 2022 generation sub-challenge consists of three stages. In the first stage, each judge individually reviews each tune and rejects it if: 1) they detect plagiarism; or 2) the rhythm is not characteristic of a reel (“If the rhythm is not close to that of a reel, or cannot be played with such a rhythm, then reject”); or 3) the mode or accidentals are not characteristic of a reel (“If the mode and accidentals used in the tune are not characteristic, and cannot be made so by transposition, reject”). In the second stage, for all tunes not rejected, each judge individually scores tunes in two qualities: Structure and Melody (1 = poor, 2, 3, 4, 5 = excellent). These two qualities were found to be the most important of the five qualities used in the 2020 challenge [1]. Each judge also elects a number of tunes for discussion in the third stage. The third stage of the challenge involves the judges working together to select tunes to recognize with awards (if any).

The score sheet provided to the judges defines the two qualities as follows (arising from expert elicitation with one of the judges):

- *Structure*: “The structure is characteristic of the reels in O’Neill’s ‘1001’ (Could you dance a reel to it? Are there 8 bars in it? Does it have the right alignment of notes within the bar?)”
- *Melody*: “The melody is characteristic of the reels in O’Neill’s ‘1001’ (Is the melody consistent with what a reel is? Does it play like a reel? Does it have a connection to Irish melody, referring to the old slow airs that predate the dance music?)”

We specifically tell the judges to consider the following in their evaluation of the tunes: “Please work individually without discussing (we will discuss after everyone has submitted scores)”; “Do not judge too harshly missing repeat signs (many reels in O’Neill’s do not specify repeats)”; “Aim to spend on average 5 minutes a tune.”

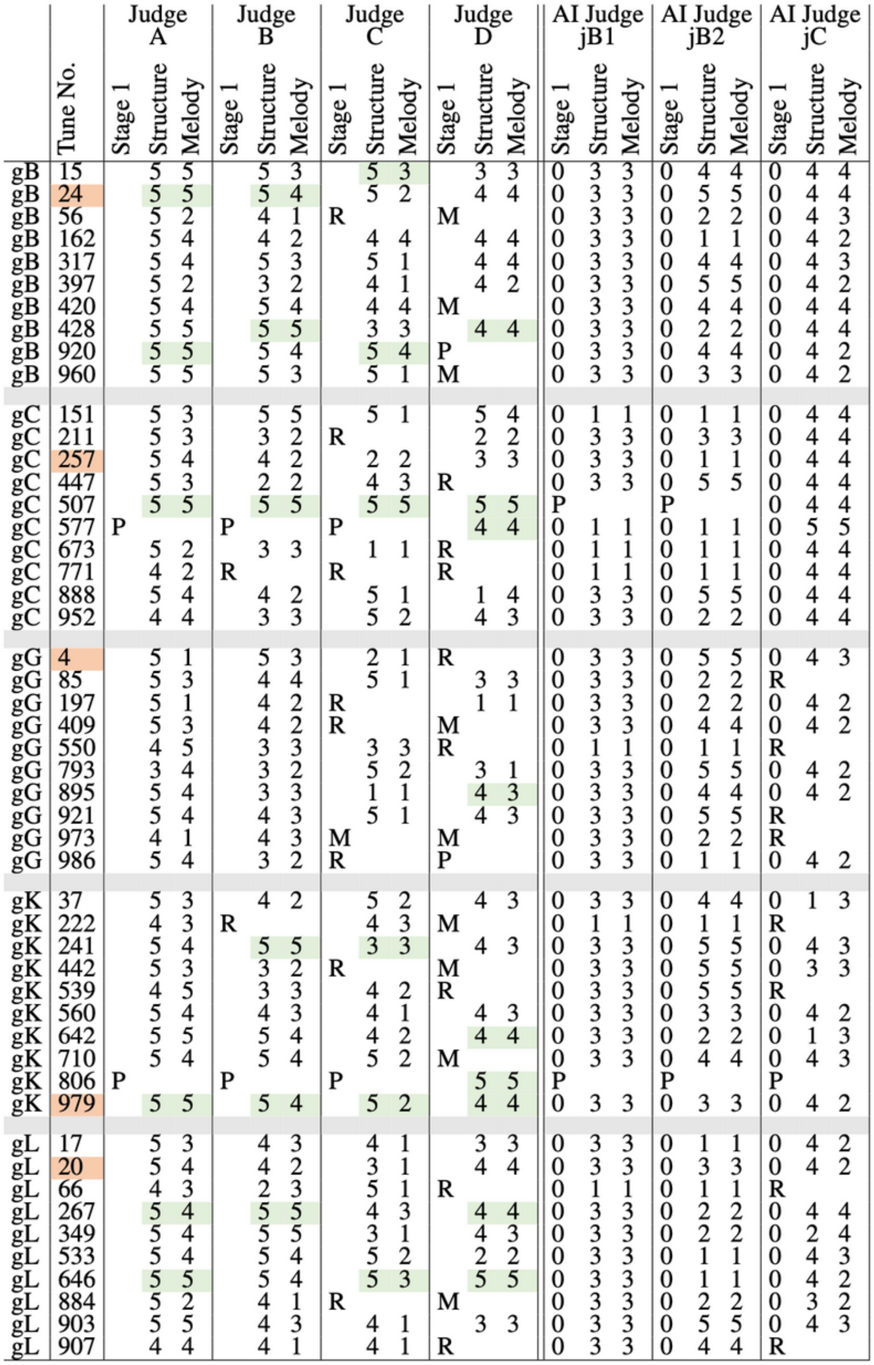
The design of the 2022 *judging* sub-challenge is as follows. Each participant receives the set of abc-notated tunes being evaluated in the generation sub-challenge, as well as the score sheet used by the human judges. The instructions are to “build an artificial judge that will imitate the four human judges” in the first two stages, i.e., rejection in stage 1, and scoring structure and melody in stage 2. The participant is to submit a file with comma-separated values with each row identifying tune number, then the result of stage 1 (*P* – reject due to plagiarism, *R* – reject due to rhythm, *M* – reject due to mode/accidentals, or *0* – no rejection), and then, if not rejected, the result of stage 2 in structure (a value 1–5) and finally the result in melody (a value 1–5). AI judges are compared with real judges in stage 1 by counting the number of coincident rejections. AI judges are compared with real judges in stage 2 by computing the mean minimum absolute difference of scores for non-rejected tunes. To make this more concrete, if the four human judges rate a tune in a quality as (2, 3, 4, 5) and an AI judge rates it a 1, then the minimum absolute difference is |2-1| = 1. These values are then averaged over the tunes passing stage 1. Two benchmark AI judges are included, both checking for plagiarism by comparing the intervalic content of a candidate tune to the 350 reels in O’Neill’s ‘1001’. One benchmark compares the metric structure (sequence of bar lines) of a non-rejected to to the metric structures in O’Neill’s. If the structure does not exist, then the structure and melody of the tune are scored with 1. If the structure does exist, then the structure and melody are scored with 3. The second benchmark just randomly selects an element from {1,2,3,4,5} and applies it to both the structure and melody of a non-rejected tune.

The design of the *tune titling* sub-challenge is as follows. Each participant receives the set of abc-notated tunes being evaluated in the generation sub-challenge. The instructions are to “build an artificial system that generates titles for given tunes”. The participant is to submit a comma-separated value file with each row containing the tune number, and the generated title. For each tune elected by any judge for discussion and recognition in the third stage of the generation sub-challenge, the organizers apply two titles from benchmark systems. The first benchmark is human titling: we create a title thought to be appropriate (without knowledge of what the other titles are). The second benchmark is naive titling with a language model: we prompt a GPT-2 model with ten titles of Irish tunes selected at random from thesession.org, and have it generate a new title. A list of possibilities titles for each tune (randomized order) is presented to the judges at the conclusion of the Stage 3 of the generation sub-challenge. The judges discuss the choices and vote for any number of them (or none of them).

	Participant (ID)	Approach

Generation	Benchmark (gB)	*folk-rnn* (v2), seed with start and 4/4 meter tokens, filter by structure and pitch range
	Clare (gC)	GRU-based language model, filter by structure
	Kerry (gK)	Variational autoencoder, filter by structure [5]
	Galway (gG)	Statistically informed recombination of material in O'Neill's, as in EMI [6]
	Limerick (gL)	as Benchmark, but with beam search ($n = 2$)
Judging	Benchmark 1 (jB1)	plagiarism detection using intervalic content; metric structure comparison: if not plagiarised, score structure and melody "1" if structure does not exist in O'Neill's, or score each "3"
	Benchmark 2 (jB2)	plagiarism detection using intervalic content; if not plagiarised, select at random from {1,2,3,4,5} and apply to both structure and melody
	Clare (jC)	plagiarism detection by 6-grams; negative log likelihood to score melody; self-similarity matrix to score structure
Titling	*Benchmark* (tB)	Prime GPT-2 with ten existing randomly selected tune titles and generate a new one
	Clare (tC)	Prime GPT-2 to create tune titles; select from these by comparing embeddings of tunes and titles using contrastive learning
	Human (tH)	Human titling of tune after listening

Table 1: Participants of the three sub-challenges of *The Ai Music Generation Challenge 2022*.



		Judge A			Judge B			Judge C			Judge D			AI Judge jB1			AI Judge jB2			AI Judge jC		
	Tune No.	Stage 1	Structure	Melody	Stage 1	Structure	Melody	Stage 1	Structure	Melody	Stage 1	Structure	Melody	Stage 1	Structure	Melody	Stage 1	Structure	Melody	Stage 1	Structure	Melody
gB	15		5	5		5	3		5	3		3	3	0	3	3	0	4	4	0	4	4
gB	24		5	5		5	4		5	2		4	4	0	3	3	0	5	5	0	4	4
gB	56		5	2		4	1	R			M			0	3	3	0	2	2	0	4	3
gB	162		5	4		4	2		4	4		4	4	0	3	3	0	1	1	0	4	2
gB	317		5	4		5	3		5	1		4	4	0	3	3	0	4	4	0	4	3
gB	397		5	2		3	2		4	1		4	2	0	3	3	0	5	5	0	4	2
gB	420		5	4		5	4		4	4	M			0	3	3	0	4	4	0	4	4
gB	428		5	5		5	5		3	3		4	4	0	3	3	0	2	2	0	4	4
gB	920		5	5		5	4		5	4	P			0	3	3	0	4	4	0	4	2
gB	960		5	5		5	3		5	1	M			0	3	3	0	3	3	0	4	2
gC	151		5	3		5	5		5	1		5	4	0	1	1	0	1	1	0	4	4
gC	211		5	3		3	2	R				2	2	0	3	3	0	3	3	0	4	4
gC	257		5	4		4	2		2	2		3	3	0	3	3	0	1	1	0	4	4
gC	447		5	3		2	2		4	3	R			0	3	3	0	5	5	0	4	4
gC	507		5	5		5	5		5	5		5	5	P			P			0	4	4
gC	577	P			P			P				4	4	0	1	1	0	1	1	0	5	5
gC	673		5	2		3	3		1	1	R			0	1	1	0	1	1	0	4	4
gC	771		4	2	R			R			R			0	1	1	0	1	1	0	4	4
gC	888		5	4		4	2		5	1		1	4	0	3	3	0	5	5	0	4	4
gC	952		4	4		3	3		5	2		4	3	0	3	3	0	2	2	0	4	4
gG	4		5	1		5	3		2	1	R			0	3	3	0	5	5	0	4	3
gG	85		5	3		4	4		5	1		3	3	0	3	3	0	2	2	R		
gG	197		5	1		4	2	R				1	1	0	3	3	0	2	2	0	4	2
gG	409		5	3		4	2	R			M			0	3	3	0	4	4	0	4	2
gG	550		4	5		3	3		3	3	R			0	1	1	0	1	1	R		
gG	793		3	4		3	2		5	2		3	1	0	3	3	0	5	5	0	4	2
gG	895		5	4		3	3		1	1		4	3	0	3	3	0	4	4	0	4	2
gG	921		5	4		4	3		5	1		4	3	0	3	3	0	5	5	R		
gG	973		4	1		4	3	M			M			0	3	3	0	2	2	R		
gG	986		5	4		3	2	R			P			0	3	3	0	1	1	0	4	2
gK	37		5	3		4	2		5	2		4	3	0	3	3	0	4	4	0	1	3
gK	222		4	3	R				4	3	M			0	1	1	0	1	1	R		
gK	241		5	4		5	5		3	3		4	3	0	3	3	0	5	5	0	4	3
gK	442		5	3		3	2	R			M			0	3	3	0	5	5	0	3	3
gK	539		4	5		3	3		4	2	R			0	3	3	0	5	5	R		
gK	560		5	4		4	3		4	1		4	3	0	3	3	0	3	3	0	4	2
gK	642		5	5		5	4		4	2		4	4	0	3	3	0	2	2	0	1	3
gK	710		5	4		5	4		5	2	M			0	3	3	0	4	4	0	4	3
gK	806	P			P			P				5	5	P			P			P		
gK	979		5	5		5	4		5	2		4	4	0	3	3	0	3	3	0	4	2
gL	17		5	3		4	3		4	1		3	3	0	3	3	0	1	1	0	4	2
gL	20		5	4		4	2		3	1		4	4	0	3	3	0	3	3	0	4	2
gL	66		4	3		2	3		5	1	R			0	1	1	0	1	1	R		
gL	267		5	4		5	5		4	3		4	4	0	3	3	0	2	2	0	4	4
gL	349		5	4		5	5		3	1		4	3	0	3	3	0	2	2	0	2	4
gL	533		5	4		5	4		5	2		2	2	0	3	3	0	1	1	0	4	3
gL	646		5	5		5	4		5	3		5	5	0	3	3	0	1	1	0	4	2
gL	884		5	2		4	1	R			M			0	3	3	0	2	2	0	3	2
gL	903		5	5		4	3		4	1		3	3	0	3	3	0	5	5	0	4	3
gL	907		4	4		4	1		4	1	R			0	3	3	0	4	4	R		

Table 2: Judge ratings of tunes generated by submitted systems summarised in Tab. 1. In Stage 1, tunes marked “M” means reject due to uncharacteristic meter; and “R” means reject due to uncharacteristic rhythm. Highlighted tune numbers (orange) are those elected by participants for evaluation. Highlighted ratings of a tune (green) denote it was singled out by the judge as a favorite.

Results of the Generation Sub-Challenge

The 2022 challenge attracted five participants in the generation sub-challenge. Table 1 summarises the submissions and their approaches. The scores resulting from Stages 1 and 2 are shown in Table 2. In total, the human judges (A, B, C, D) evaluated 50 transcriptions. Two of these were rejected in Stage 1 due to plagiarism: 577 by gC (original named “The Dairy Maid") and 806 by gK (“Green Garters”). Several tunes were rejected in Stage 1 by particular judges for failing the rhythm criterion (R) and mode criterion (M). Judges C and D were the most sensitive to these two criteria, where judge C rejected 8 tunes due to rhythm and one due to mode; and judge D rejected eight tunes due to rhythm and nine due to mode. At the conclusion of Stage 2, each judge elected a set of tunes for discussion in Stage 3: judge A elected five (24, 507, 646, 920, 979); judge B elected two (241, 267); judge C elected five (51, 241, 507, 646, 920); judge D elected five that were not plagiarised (428, 507, 642, 646, 895). Judge B had misunderstood the instructions as to select only two. During discussions in Stage 3, judges A, B and C expanded their elections to six, and judge D to seven (all elections highlighted green in Table 2). Eleven tunes were considered by the judges for recognition during Stage 3: 24, 51, 241, 267, 428, 507, 642, 646, 895, 920, and 979. Each submitted collection contains at least one tune receiving at least one vote. The non-plagiarized tunes generated by the benchmark garnered seven votes over four tunes; those of gK received seven votes over three tunes; those gL received six votes over two tunes; those of gC received four votes on one tune; and finally those of gG received one vote on one tune. The total time spent in Stage 2 by the judges were: A spent 4 hours, B and C spent 6 hours, and D spent 7 hours.

The discussion in Stage 3 between all four judges and the organiser lasted about 57 minutes. Tune 986 was discussed as possible plagiarism, but in the end decided to be original. Tune 920 was also discussed as possible plagiarism, with judge A and C identifying its first part as “kind of ‘The Abbey’ reel, but 90% ‘The Abbey’ reel”, but judge B argued that, “There’s lots of tunes that can be quite similar; it would be harsh to put it down as totally plagiarised I suppose.” Judge C mentioned that the second part could be a third part to “The Abbey” reel. During the conversation it became apparent that judges A and B were allowing minor changes to tunes in their assessment, while judges C and D were more strict. Judge B said, “I supposed the angle I was coming from was when we chose the winning jigs in 2020 we made some minor adjustments. Lots of these are going to come out with a fault or two and I remember there was one note in the jig that we chose last time that was truly awful; and we changed it and it made a job of the tune.”

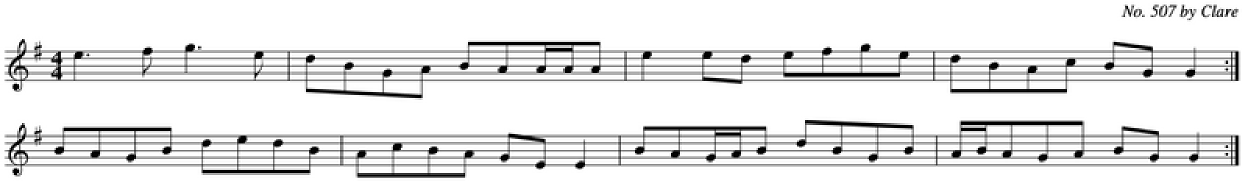
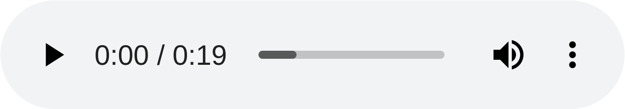


Figure 2: This tune generated by Clare is awarded first prize in the generation sub-challenge.



The synthesis of tune 507 (gC) heard by the judges.

From the comments made in Stage 2 and 3, the clear favorite is tune 507 (gC), notated in Fig. 2. Judge A wrote on their score sheet: “Excellent! Everything right! That’s a 5+. Not plagiarised as far as I can see (I suspected it might be since it was so good).” Judge B wrote, “Very simple reel, but adheres to all the variables regarding reel structure. Great rhythm. Makes good use of repetition. Very simple but consistent and phrases are easy to remember. Excellent simple reel that sits well with the tradition.” Judge C wrote, “Nice tune”; and judge D wrote: “easy to remember, easy to follow and catch. Uplifting, bright and fun to dance to.” Judge B was the only one to not elect this tune, and remarked in the Stage 3 discussion: “It was a marginal call for me not to pick it. In a different mood on a different day [I could have elected it], so I would be quite happy to give it my vote as well. I suppose compared to some of the reels I picked, I didn’t think the melody was as deep; but it’s a bonafide reel, and a nice simple tune. So I wouldn’t object to it getting the prize at all.” All judges concurred that this tune was far and away the best of the lot.

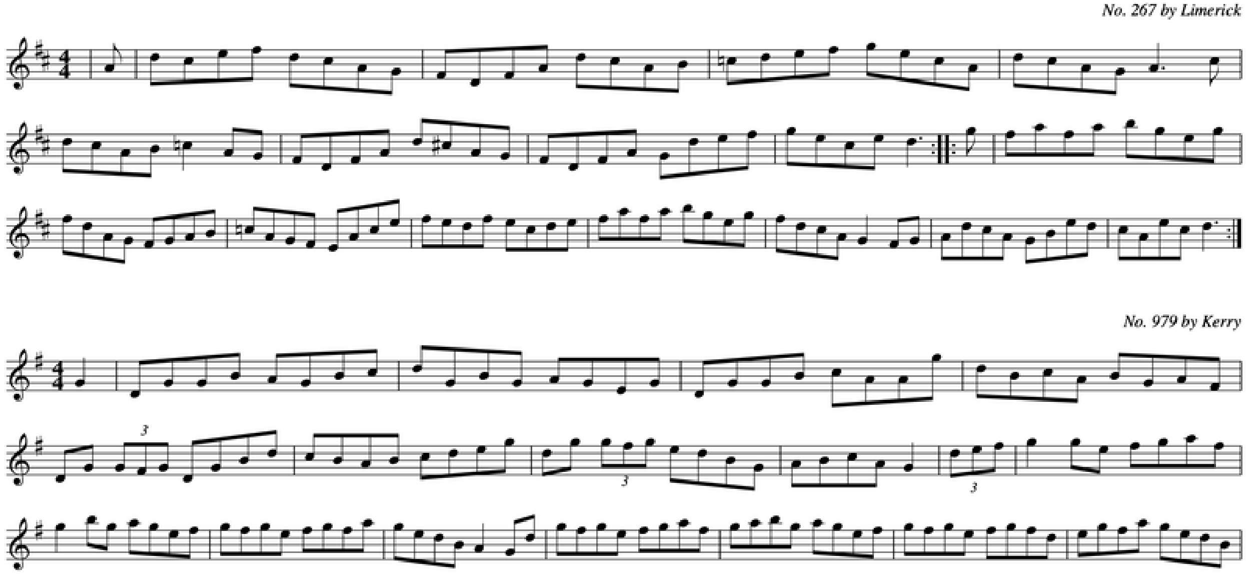
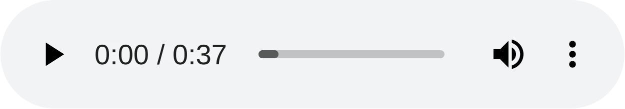
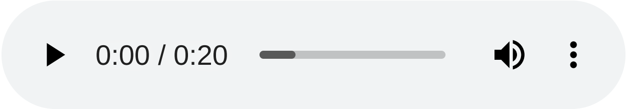


Figure 3: These two tunes by two different systems share second prize in the generation sub-challenge.



The synthesis of tune 267 (gL) heard by the judges.



The synthesis of tune 979 (gK) heard by the judges.

Discussion in Stage 3 converged on awarding two tunes second prize: 267 (gL) and 979 (gK), each notated in Fig. 3. For tune 979, judge A wrote on their score sheet: "Good tune". Judge B wrote, "Structurally sound in every way. Easy to play and always has the feel of a reel when played. Good melody. Good use of repetition. Plays well on accordion and worth learning. Phrases are strong and stand out. Melody definitely belongs in the tradition." Judge C wrote, “First part stronger.” Judge D wrote that it is not really danceable. In the Stage 3 discussion, the judges remarked on their confusion listening to the mp3 synthesis due to the treatment of the pickup. After we auditioned the tune again with the pickup removed, judge D remarked, “That pickup really put me off and so I wrote that it is not danceable. But I really like this tune now, and would put it very high on my list. I like the balance between the first and second part.” Judge C remarked, “In fairness, it’s not a bad tune. Maybe eliminating that [pickup] has put a different complexion on it. Maybe also when you played it slower here than in the mp3 it sounded a lot nicer. But compared to 507, tune 979 is miles behind it.” Judge B remarked, “It’s a good tune. I have nothing negative to say about it.”

For tune 267 (gL), judge A wrote on their score sheet: “ok. Bar 11 2nd beat strange. Interesting. Nice changes c#/c”. Judge B wrote, “Perfect in structure. Great natural rhythms. Beautiful melody – Hopefully it’s new (I’m unfamiliar with the tune if it already exists.); a fantastic tune. A tune I’d be happy to play. Well done!" Judge C wrote: “First part needs work. Nice minor.” Judge D wrote: “The melody doesn’t go anywhere. There is no climax.” In the Stage 3 discussion, judge B said that it is their favorite tune of the 50: “It was the only melody that I played that overall had kind of a deeper melody – something that, say, one of the greater tune writers like Finbarr Dwyer may have wrote. I found it really, really interesting, especially on the accordion. Maybe it is an accordion thing. But listening on the computer does not do it justice. The more I played it the more I felt I could do something with it. Especially going from major to minor. The tune managed to endure those key changes. ... The other thing I like is that the melody progressed in a very logical way. It’s got a lot of characteristics of tune that was written by a very good tune writer, like a highly respected writer of tunes within the tradition. Not somebody who writes Mickey Mouse tunes.” Judge A and D agreed with what judge B said, that it is a good tune, and ended up voting for it. Judge C remarked, “I would absolutely not vote for this tune. It’s well-structured, well put together, what have you, but it’s not a tune I would bother playing. Going from the major to the minor, what have you, just doesn’t do it for me. I don’t want to say it’s wrong, or gimmicky, whatever, it just doesn’t do it for me." During the Stage 3 discussion, the decision was made that this tune share second place with 979, and that tunes 507, 267 and 979 are the three best tunes of the 50.

Several tunes were elected by several judges, but in the end did not receive awards for a variety of reasons. Four of these are notated in Fig. 4. Reel 646 (gL) received three votes. Judge D wrote in their score sheet that the structure is “perfect”; and of the melody: “perfect, interesting and inviting into a mystical magical world. I really like the bar 2 and 7, they do the job, they give just a little more than the bar before which I think is a great way to build up a tune!” Judge C wrote, “Not a bad tune. Second part needs work.” Judge A wrote, “Good tune. 1st part slightly repetitive, but 2nd part interesting.” Judge B wrote, “Structurally sound reel, reads and plays well. Good development of melody with ample use of repetition in 1st part to allow phrases to be memorable. 2nd part slightly less so but it’s just a matter of taste. Another musician may like it more than myself. Excellent effort that many traditional players would be happy to play.” In the Stage 3 discussion, judge B said, “It was the second part I didn’t like. It was jumping up and down too much... just a touch of James Last Orchestra. I didn’t like it. ... One of the criteria I put the tunes: If was down at the Willie Clancy festival playing in a bar, would I play that in front of my peers, and no, I wouldn’t play it. I love the first part, but I could never ever see myself playing the second part – not even to save somebody’s life.” Judge C agreed and said, “The first part is very good, but the second part is a bit gimmicky. ... Like a piano accordion, Scottish type thing. ... I do think that with a couple of adjustments of a few of the bars in the second part it would not be a bad tune. But it’s not as good as 507.” Judge A remarked that some tunes from Donegal have similar features. Judge B offered further observations: “You know what it sounds like. It sounds like when the fight breaks out in the film *The Quiet Man*, and they are all just beating each other up. And that’s what’s going on in the background.” Judge D remarked: “The second part sounds modern. The first part I really like. The second part is a little embarrassing somehow.”

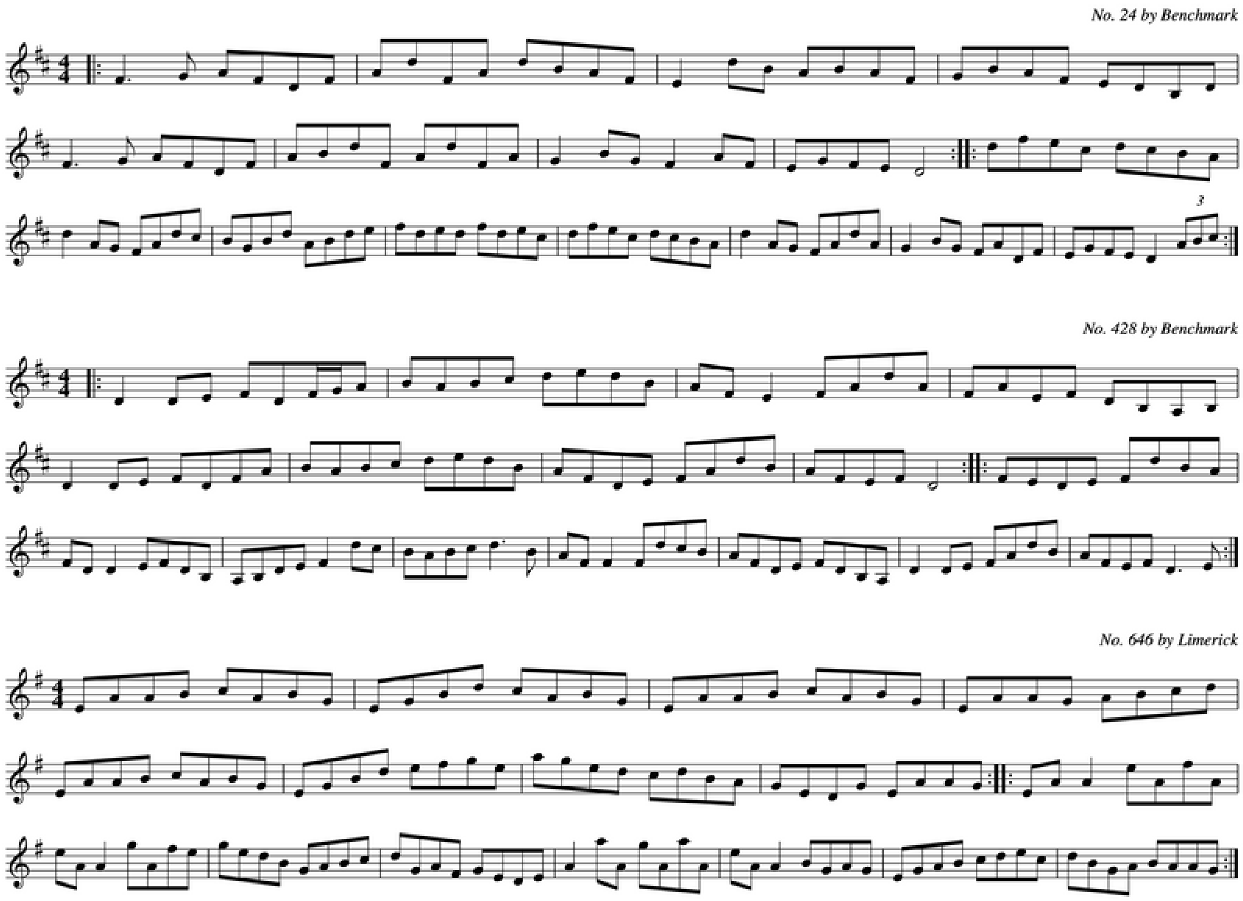
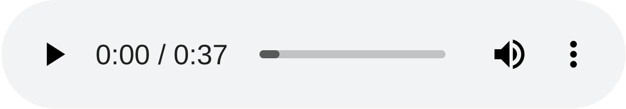
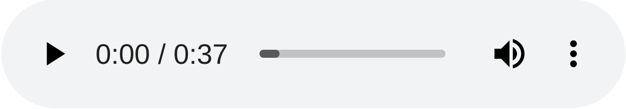


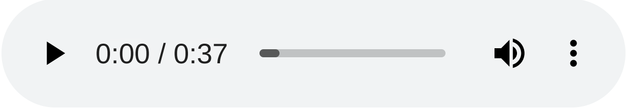
Figure 4: These three tunes were elected by several judges, but not awarded prizes.



The synthesis of tune 24 (gB) heard by the judges.



The synthesis of tune 428 (gB) heard by the judges.



The synthesis of tune 646 (gL) heard by the judges.

Finally, reels 24 (gB) and 428 (gB) each got two votes. Judge A wrote on their score sheet, “ok. Typical structure. 2nd part higher. 1st part ending comes back in 2nd part.” Judge B wrote, “Perfect structure. This is very consistent with reels in ‘1001’, with good use of repetition in the melody. Good use longer notes in a consistent fashion gives the tune the feel of an Irish reel. Excellent attempt.” Judge C wrote, “Not a bad attempt, needs work.” In Stage 3, judge B remarked, “I have nothing negative to say about this tune at all. I did like it. If everyone was voting for it, I would have no reason to say ‘no.’ But there is no reason to put it above 507.” Judge C said, “The first part doesn’t do it for me. The second part is a lot better.” Judge D said the same, and “The first part is messy and I get tired of the melody... but I would like to hear this on the [accordion], but not on the fiddle.” Judge A said, “If I would play this tune, I would adjust it a bit. But that’s the usual thing with these tunes. Even O’Neill’s.” Judge A mentioned that it was not among their top tunes, but when reminded that they had elected it as a favorite they said, “Oh I did! Well not today.” Judge B remarked that they feel the same with many of these tunes: “There’s certain brilliant classic tunes and you play them, every time they just warm you up. They’re brilliant all the time. There’s something just a little bit short with each of them, that depending on what mood you’re in on the day you may like it you may not. There’s just a fine line there with all of them.”

For reel 428 (gB), judge A wrote on their score sheet, “ok. Better when transposed to G.” Judge B wrote, “Excellent in structure. Very rhythmical and suited to dancing. Excellent melody very much in keeping with ‘1001’. A few small suggested edits (most notably 1st part bar 3 AF#DE instead of AF#E–). Excellent tune.” Judge C wrote, “An ok tune. Not very musical.” Judge D wrote, “A little boring but ok. Cool B part.” In Stage 3 Judge B said, “I could have just as easily put this in my list, but I felt I just had to narrow my list because I had too many. I like this tune, and I have no hesitation voting for it. It’s a good effort.” Judge A said, “There’s nothing really wrong with it. It’s just a bit bland I would say. That’s the only thing. I’ve heard worse tunes played in sessions.” Judge D agreed.

Results of the AI Judging Sub-Challenge

The last three columns of Table 1 show the responses of the two benchmark AI judges (jB1, jB2), and the submitted system (jC). Each AI judge detected only one of the two plagiarized tunes, giving each a plagiarism true positive and plagiarism false negative of 1. Both benchmark judges have 1 plagiarism false positive, while jC has none. jB1 and jB2 were not designed to reject based on meter or accidentals, so their rhythm true positives (RTP) and rhythm false positives (RFP) are zero. jC however has 5 RTPs and 3 RFPs. Clearly for Stage 1, jC is the more sensitive AI judge of the three.

For Stage 2, we compute for each of the two qualities the minimum absolute error between AI judge scores and the corresponding human judge scores. For the 47 of 50 tunes passing its Stage 1, jB1 has a mean minimum absolute error of 0.77 in Structure. For Melody, this is 0.43. These are all larger for Benchmark 2: jB1 has a mean minimum absolute error of 0.92 and 0.66 for structure and melody, respectively. For the 40 of 50

tunes passing its Stage 1, the mean minimum absolute error is 0.53 and 0.45 for structure and melody, respectively. Clearly for Stage 2, jC is the best judge.

Results of the Tune Titling Sub-Challenge

After concluding the Stage 3 discussion about which tunes to recognise with awards, the human judges were tasked with commenting on three titles for each of the eleven tunes nominated in Stage 2. The judges found this exercise very entertaining, but were reluctant at times to vote for any of the options. It took about 15 minutes to complete this activity. In summary: titles generated by the human received 22 votes; those generated by tC received one vote; and those generated by tB received zero votes. Several generated titles received votes against, as being "weird" or hard to interpret, plagiarised, or as having problematic connotations. Some of these are discussed below.

The titles given to 507 (gC) were: "And the rest they all went" (tB), "Lad O'Beirne's" (tC), and "A winter in [redacted]" (tH) – the redacted portion is a place in Ireland. Upon seeing these, one of the Irish judges said, "Oh Jesus, no. [redacted] evokes all sorts of images here in Ireland. It has a certain connotation here. It's a dreary place with notoriously mean people." The other Irish judge said, "If you lived in Ireland, you'd rather a stretch in the Gulag than in [redacted]." One of the Swedish judges, however, liked this title very much: "That title has 10 points from me, and the others have like two points. I don't have many associations with [redacted]. I'm sorry."[2] The Irish judges agreed it would be insulting to name this winning tune "A winter in [redacted]". Furthermore, the Irish judges mentioned that Lad O'Beirne is a noted Irish fiddler, and it would not be right to name the tune after him when he did not compose it – though there are several tunes called "Lad O'Beirne's". The other title was agreed to be "weird".

The titles given 24 (gB) were: "The Pogo Tree" (tB), "The Changelings" (tC), and "Money in the Pocket" (tH). One of the Swedish judges asked whether "The Changelings" means anything, and one of the Irish judges remarked that "it is not a nice thing in Ireland. It's basically when people had handicapped kids in Ireland long ago, they said the fairies swapped them for the healthy ones. I'd scrub this title." Tune 642 (gK) was given the titles: "The Orange Brothers' Liberty Club" (tB), "O'Connell meets the Queen" (tH), and "The Gosson That Bate His Father" (tC). The last title is plagiarised. A Swedish judge asked what "orange" means in the first, and both Irish judges remarked, "You don't want to know." In this case, the connotation is of the political and religious troubles Ireland has faced.

Conclusion

While participation in *The Ai Music Generation Challenge 2022* was not very high, it presented a rich variety of outcomes. First, all judges remarked in the Stage 3 discussion that there was a "incredible" increase in quality as compared to the 2020 challenge. Judge A remarked, "Last time we had either plagiarism or weird stuff. Here we have loads of plausible tunes." Judge B remarked, "There's a lot of them that are reasonably good tunes. Everything had to be listened to and everything had to be played – but none of them were absolutely mind blowing.” Second, the discussions between the judges shows how their assessments are shaped by one’s instrument, the relationships of the tradition to extra-musical factors such as films, variety bands, geography, and the presentation of the materials. One cannot just evaluate a melody based on its tonal characteristics, but also how it fits on a given instrument, and within some practice.

Judge D remarked on trouble understanding some tunes in the beginning because the pickup in the MIDI synthesis is not played appropriately. Judge D had to read the notes to understand what was going on in the synthesis. This motivates synthesizing the evaluated material more carefully such that each is presented as musically as possible. Each tune could be synthesized as slow and fast renditions. Judge D also remarked in Stage 3 that if they “had time to learn all the tunes, they would have given them different scores.” Evaluation with respect to the 350 reels in O’Neill’s is anything but straightforward. The scores given to the submitted tunes only serve as a guide for discussion. The positive and negative characteristics of each tune come out in the discussion between expert practitioners, demonstrating levels of ambiguity and inconsistency when it comes to evaluation – both between and within judges. The anchoring of The Ai Music Generation Challenge to a very specific and well-defined kind of practice helps limit the ambiguity of evaluation.

Finally, of the three sub-challenges, it appears that the tune titling one is by far the hardest. The outcomes of this sub-challenge reveals very intriguing aspects of the socio-cultural contexts in which this music is practiced. Titling tunes in ways that are divorced from these contexts (via human or machine) can produce results that are entertainingly obtuse, or perhaps inappropriate or offensive. How to appropriately link a piece of music with a descriptive and inoffensive – or non-weird – text seems to be a problem naturally addressable with multimodal models.

Acknowledgments

This work was supported by the grant ERC-2019-COG No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism.

Footnotes

1. The human judges in 2022 are the same as for the 2020 challenge: Jennikel Andersson, Kevin Glackin, Henrik Norbeck, and Paudie O’Connor. ↩

2. I was unaware of these associations with [redacted] when I created the title. ↩

References

- Amerotti, M. (2022). *Latent representations for traditional music analysis and generation* [Mathesis]. University of Bologna, Italy. ↩
- Cope, D. (1996). *Experiments in Musical Intelligence*. A-R Editions. ↩
- https://github.com/boblsturm/aimusicgenerationchallenge2022

↩

- O’Neill, F. (1907). *The Dance Music of Ireland: O’Neill’s 1001*. Chicago. ↩
- Sturm, B. L. T. (2022). The Ai Music Generation Challenge 2021: Summary and Results. *Proc. AIMC*. ↩
- Sturm, B. L. T., & Maruri-Aguilar, H. (2021). The Ai Music Generation Challenge 2020: Double Jigs in the Style of O’Neill’s “1001.” *Journal of Creative Music Systems*. ↩