An Analysis of Singing Accuracy towards Quantifying the Melodic Singability

Minami Kawahara* and Tetsuro Kitahara†
* Nihon University, Japan
E-mail: minami@kthrlab.jp
† Nihon University, Japan
E-mail: kitahara@kthrlab.jp

***Abstract*—This study is part of an effort to define "singability" by analyzing human singing voice, with the goal of enabling AI to consider this when generating vocal melodies. Specifically, we examined the pitch accuracy of 91 participants singing 18 types of melodies that combined different degrees of melodic leaps (three types), beats per minute (BPM) (three types), and lyrics (two types). As evaluation metrics, we used the Median Absolute Error ($MedAE$) between the singing voice and the reference melody, and the fluctuation of pitch within each note ($StdDev_{cent}$). The results showed that an increase in melodic leap size significantly worsened both $MedAE$ and $StdDev_{cent}$, making it the most influential factor. The effect of BPM differed; $StdDev_{cent}$ consistently decreased (indicating improved stability) as BPM increased, suggesting a familiarity effect with the melody. In contrast, $MedAE$ varied in a complex manner with BPM and leap size, indicating an increased processing load at faster tempos and the influence of familiarity. The effect of lyric type was limited compared to the other two factors.**

I. Introduction

In recent years, the automatic generation of singing melodies has garnered significant interest within the field of AI-based music generation. In the modern fields of music production and entertainment, technologies that utilize AI to generate high-quality musical pieces and singing voices are advancing, with their results beginning to see widespread adoption as virtual singers and AI VOCALOIDs. However, while it is crucial that these singing melody generation technologies are designed with human singing, they are currently limited to using vast amounts of existing musical data for training. Consequently, it is not yet clear whether the generated singing is "possible for humans to sing".

The element of "singability" is closely related not only to whether a song's melody is beautiful but also to physical and psychological aspects, such as human vocal range, phonatory characteristics, and rhythmic processing during singing. However, existing automatic generation technologies do not sufficiently consider this aspect of singability, and the generated melodies or singing data are not necessarily something that humans can sing naturally or comfortably.

The purpose of this research is to attempt to define "singability", which is the feasibility of a person being able to sing a melody, by analyzing the actual voices sung by experiment participants. In this paper, as a preliminary step, we collect fundamental frequency from actual singing voices and verify the accuracy of that singing.

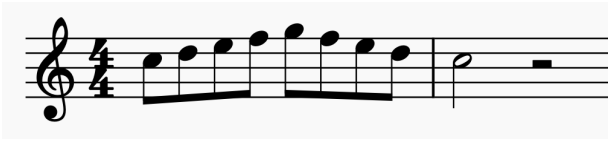
A previous study [1] focusing on singability calculated song difficulty based on subjective evaluations. While that research did not delve deeply into the accuracy or physiological aspects of actual singing, factors, such as melodic leaps, melodic complexity, and human vocal range are broadly considered to influence singability. Therefore, in this analysis, we will investigate melodic leaps, along with differences in BPM and lyrics (articulation).

II. Related Work

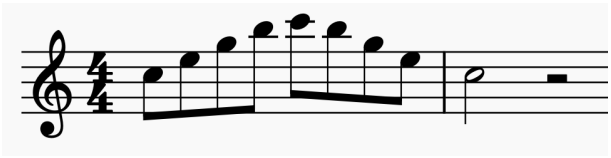
In related work, a preceding study on a music search system focusing on singability [1] should be mentioned. While this research calculates the difficulty of songs based on subjective evaluations, it does not delve deeply into the accuracy or physiological aspects of the actual act of singing. Although many studies touch upon singability, the majority consider it in the context of translation [2][3]. Some studies aim to improve singability through practice and instruction [4], while others examine the singability of existing songs [5], but these analyses were not targeted at individuals with little singing experience. Furthermore, research considering song difficulty based on voice speed and rhythmic irregularity [6] exists, but it does not focus on singability. While some studies on singing generation AI consider singability [7][8], these were limited to ensuring consistency between lyrics and melody and aligning dynamics and note durations.

In this study, forming a part of a broader effort to define "singability", aims to objectively measure singing accuracy through experiments involving human participants and thereby identify the factors influencing it. As a preliminary step, this paper presents a detailed analysis of a singing experiment. This experiment involved 100 participants singing 18 types of melodies, which were systematically constructed from combinations of distinct melodic patterns (such as stepwise and leaping progressions), various Beats Per Minute (BPM) levels (80, 120, and 150), and different lyric types (two types). The analysis focused on evaluating the pitch accuracy of the collected singing data, derived from estimated fundamental frequency values.

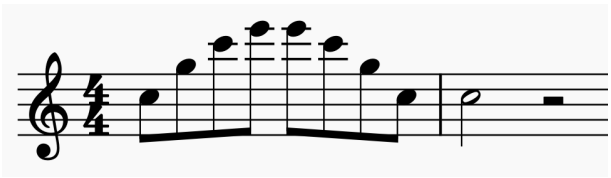
979-8-3315-7206-8/25/$31.00



(Phrase 1) Melodic Pattern 1



(Phrase 2) Melodic Pattern 2



(Phrase 3) Melodic Pattern 3

Fig. 1. Singing phrases used in the experiment

III. Experiment

A. Data Used

In this experiment, participants were asked to sing along to a specified melody prepared for this study. The experimental conditions included three melodic patterns (1, 2, and 3), two sets of lyrics (lyrics 1: "dadadadadadadada," lyrics 2: "dabidabadabidabada"), and tempos of 80, 120, and 150 BPM.

Participants were asked to sing each of the 18 melodies, which were combinations of the three types of leap progressions, two types of lyrics, and three types of BPM. Examples of the melodies used for each leap type are shown in Figure 1.

The 18 melodies are classified from Sound Source 1-1-1 to Sound Source 3-2-3. In this classification, the first digit represents Melodic Patterns 1, 2, and 3, the middle digit represents Lyrics 1 and 2; and the last digit represents BPMs 80, 120, and 150.

Hereafter, the audio data recorded by the participants will be referred to as "recorded data," and the singing melodies prepared for this study will be referred to as "reference data."

For the reference data, we used the voice of the machine-generated voice, "Kasane Teto" from the voice synthesis software, Synthesizer V[1], to sing as shown in Figure 1.

B. Participants

One hundred participants, recruited through the crowdsourcing site "Lancers," took part in this study.

Participants were surveyed about their past singing experience, such as being in a choir or taking vocal lessons. The results are shown in the Table I.

[1]https://dreamtonics.com/ja/synthesizerv/

TABLE I
Singing Experiencet

Singing Experiencet	Number of people
None	87
Less than 6 months	5
6 months to less than 1 year	4
1 to less than 3 years	1
3 to less than 6 years	1
6 years or more	2

C. Experimental Procedure

1) Experiment Form: For this experiment, a form equipped with functions for playing the singing melody and recording was created using Google Apps Script. Participants sequentially recorded their singing using the recording function of this form.

2) Experimental Melodies: In the audio presented to the participants, a machine voice first sings the melody as a model. Next, to help them grasp the pitch, a melody played on an electric piano from the music production software "GarageBand" was added. Participants were asked to sing along with the model voice during the part in which the electric piano was playing.

Furthermore, for each melody, the machine voice and the electric piano were repeated alternately three times. Therefore, participants were required to sing the same phrase three times. The recording for the same melody was done in a single take unless a significant recording error occurred.

During recording, participants were required to wear earphones or headphones to listen to the model track, so it is assumed that generally only the participant's voice was recorded.

IV. Analysis Method

In this study, we analyzed the overall fundamental frequency of the music using the pYIN fundamental frequency estimation algorithm from the music analysis module LibROSA.

To analyze singability, we conducted a numerical evaluation from two perspectives: "the overall pitch accuracy of the singing data" and "the pitch accuracy for each individual note."

The necessary processing steps for the analysis are listed below in order.

A. Time Synchronization of Recorded and Reference Data via Cross-Correlation

To calculate the overall and per-note pitch accuracy, the time axes of the recorded data and the reference data must be precisely aligned. For this purpose, we calculated the cross-correlation between the recorded data and the reference data using the correlate function of the NumPy extension module.

B. Noise Reduction

Since this experiment adopts a format where participants record in their own environments, the possibility of noise contamination during recording had to be considered. To address this issue, we normalized the amplitude values of the recorded data so that the maximum value was 1.0, and

implemented a filtering process to exclude frames with an amplitude of 0.01 or less from the recording.

C. Unit Conversion

When estimating frequency, it is easier to grasp pitch differences by thinking in cents rather than Hz. Since the numerical unit for pYIN fundamental frequency estimation is Hz, we convert it to cents. The conversion formula is as follows. In this formula, f is the frequency to be converted, and f_{ref} is the reference frequency (440Hz is used in this experiment).

$$f_{cent} = 1200 \cdot \log_2 \frac{f}{f_{ref}} + 5700$$

D. Octave Adjustment

When comparing recorded data and reference data, differences in pitch from the reference data may arise due to each participant's vocal range, which can cause large fluctuations in the analysis results. However, since this fluctuation may also simply be due to an inability to hit the correct pitch, we performed a process to align pitches that were off by an octave with the reference data for a more accurate analysis. Specifically, we referenced data with +2400, +1200, 0, -1200, and -2400 cents added to the cent value of the reference data, calculated the error for each, and used the data with the smallest error for analysis.

E. Overall Pitch Accuracy of Singing Data

To evaluate the accuracy of the entire piece of music, we focused on its fundamental frequency. We used the "Median Absolute Error ($MedAE$) of the fundamental frequency in cents for the entire singing data" to evaluate the error against the reference data, based on the estimated fundamental frequency values. The calculation formula is $med(|x_i - p_i|)$.

We define two variables: x_i is the estimated fundamental frequency of the recorded data at frame i, and p_i is the estimated fundamental frequency of the reference data at frame i.

A smaller $MedAE$ indicates that the estimated result is closer to the reference data, signifying higher accuracy.

F. Pitch Stability within a Single Note

To examine the stability of pitch within a single note (i.e., whether the pitch is wavering), we divided the calculation into eighth-note intervals and computed the variance of the frequency within each interval. Then, we found the average of the variances across all intervals and took the square root of that average value. This value is defined as pitch instability (hereafter, $StdDev_{cent}$).

The formula is as follows:

$$StdDev_{cent} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{N_j}\sum_{i=1}^{N_j}(x_{ij} - \bar{x_j})^2}$$

We define x_{ij} as the estimated fundamental frequency at frame i of note j. The variable N_j is the number of frames in note j, and x_j is the average of x_ij over these frames. The total number of notes in the data is n.

A smaller $StdDev_{cent}$ value is evaluated as being more stable and well-sung.

V. Experimental Results

We present the data from 91 out of 100 participants, excluding 9 individuals from whom valid numerical values could not be obtained using the pYIN fundamental frequency estimation algorithm.

Figure 2 displays box plots of the distribution of the $MedAE$ between the recorded data and the reference data in terms of fundamental frequency in cents for each melody. In this figure, the horizontal axis represents the 18 types of melodies (3 leap types x 2 lyric types x 3 BPM types), and the vertical axis indicates the magnitude of the $MedAE$.

Figure 3 displays box plots of the distribution of the $StdDev_{cent}$ within eighth-note intervals of the recorded data's fundamental frequency in cents, for each melody. In this figure, the horizontal axis represents the 18 types of melodies (3 leap types, 2 lyric types, and 3 BPM types), and the vertical axis indicates the magnitude of the $StdDev_{cent}$.

A. Leaps

As the degree of melodic leap increased, the first quartile, median, and third quartile of both $StdDev_{cent}$ and $MedAE$ tended to rise. Specifically, for $StdDev_{cent}$, the values progressively increased from the Melodic Pattern 1 condition group to the Melodic Pattern 2, and further to the Melodic Pattern 3 condition group. Similarly for $MedAE$, the values were low (small error) in the Melodic Pattern 1 group, higher (larger error) in the Melodic Pattern 2 group, and markedly higher (very large error) in the Melodic Pattern 3 group. For example, comparing the third quartile $MedAE$ under the Lyrics 1, BPM 80 condition, the values were approximately 110 cents for 1-1-1, 180 cents for 2-1-1, and 260 cents for 3-1-1.

B. BPM

The effect of BPM on $StdDev_{cent}$ and $MedAE$ varied depending on the conditions. The first quartile, median, and third quartile of $StdDev_{cent}$ showed a stepwise decrease as the BPM increased from 80 to 150 under almost all conditions.

Regarding $MedAE$, the following trends were observed.

- In the Melodic Pattern 1 condition group, comparing 1-1-1 with 1-1-2, and 1-2-1 with 1-2-2 in Figure 2, it can be seen that the third quartile increased by about 100 cents between BPM 80 and BPM 120. On the other hand, comparing 1-1-2 with 1-1-3, and 1-2-2 with 1-2-3, it decreased by about 70 cents between BPM 120 and BPM 150.
- In the Melodic Pattern 2 condition group, comparing 2-1-1 to 2-1-3, and 2-2-1 to 2-2-3 in Figure 2, the median can be seen to increase as BPM goes from 80 to 150. Also, from 2-2-1 to 2-2-3, the third quartile can be observed to increase.

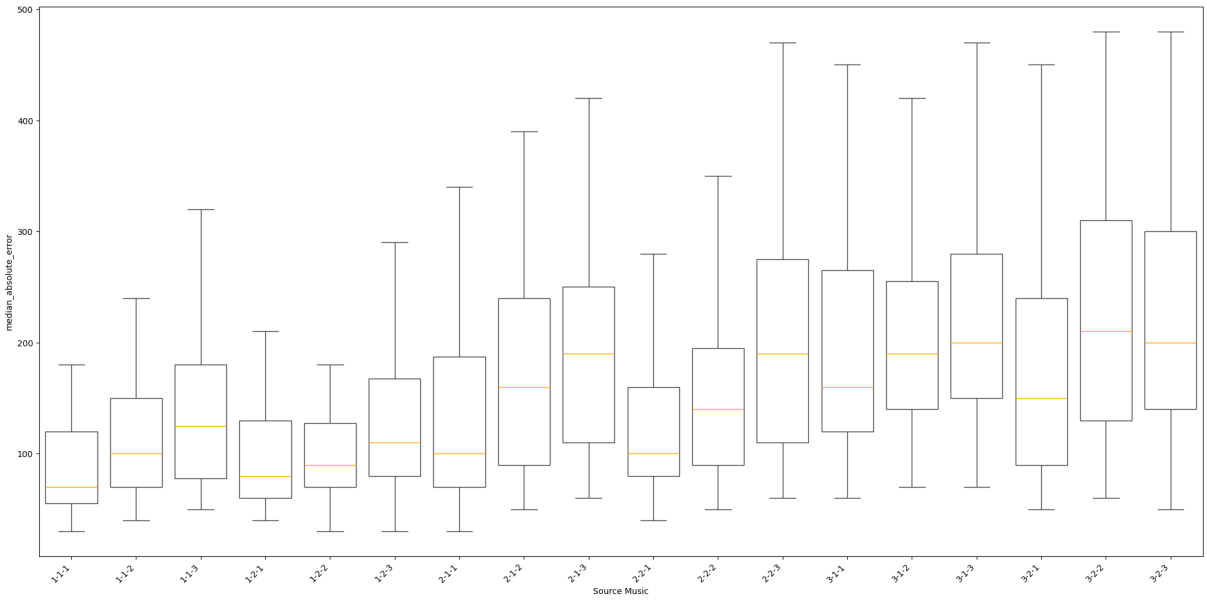


Fig. 2. Distribution of $MedAE$ between recorded and reference data for each melody (x-y-z; x = melodic leap, y = lyrics, z = BPM)

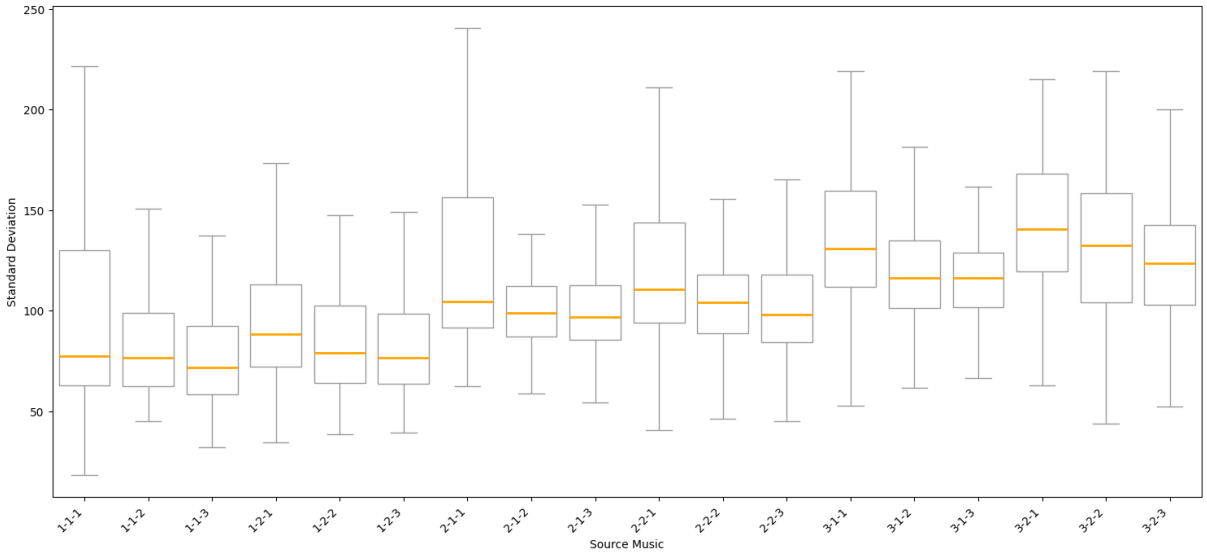


Fig. 3. Distribution of $StdDev_{cent}$ of recorded data for each melody (x-y-z; x = melodic leap, y = lyrics, z = BPM)

- In the Melodic Pattern 3 condition group, the conditions were divided by lyrics.

  Lyrics 1 In Figure 2, from 3-1-1 to 3-1-3, when comparing BPM 80 and BPM 120, the median remains unchanged, while from BPM 120 to BPM 150, the median increases and the third quartile decreases.

  Lyrics 2 In Figure 2, from 3-2-1 to 3-2-3, when comparing BPM 80 and BPM 120, the median and third quartile increases by about 150 cents, and then decreases from BPM 120 to BPM 150.

C. Lyrics

The influence of lyrics on $StdDev_{cent}$ and $MedAE$ was small compared to the effects of the degree of leap and BPM.

Particularly notable differences include the decrease in the third quartile of $StdDev_{cent}$ between 1-1-1 and 1-2-1, and 2-1-1 and 2-2-1, and the increase in the median of $MedAE$ between 3-1-2 and 3-2-2.

VI. Discussion

This section, based on the results presented in Chapter V, will address the effects of the degree of leap, tempo (BPM),

and lyrics on pitch stability and accuracy during singing, as well as the relationships between these metrics.

A. Leaps

As shown in Section V-A, the degree of leap had the a strong negative impact on both pitch stability and accuracy. This is thought to be because large leaps demand more precise pitch sense (adjustment via auditory feedback of the target pitch) and vocal cord control (rapid and accurate attainment and maintenance of the target pitch) from the singer. The increase in leap width increases these cognitive and motor loads, which can be interpreted as the root cause of both the fluctuation of pitch within a note (decrease in stability) and the deviation of pitch from the reference (decrease in accuracy).

B. BPM

The influence of BPM presented a different picture for stability and accuracy.

1) Pitch Accuracy: Regarding MedAE, it can be observed that the first quartile, median, and third quartile generally showed a monotonic increase across almost all conditions. In contrast, in conditions such as 3-1-2 and 3-2-3, it was sporadically observed that the median and third quartile errors were lower than at slower tempi. This is considered to be due to participants who initially struggled with singing becoming accustomed to the melody as the experiment progressed sequentially.

2) Pitch Stability: Regarding $StdDev_{cent}$, it is clear that for almost all conditions, the first quartile, median, and third quartile of $StdDev_{cent}$ decrease as the BPM increases from 80 to 150. This suggests that familiarity with the melody has a significant influence on $StdDev_{cent}$.

3) Comprehensive Tempo Effects on Pitch Control: As tempo increased, the study indicated a dichotomy: while achieving pitch stability tended to become easier, maintaining precise pitch accuracy did not necessarily improve commensurately. This can be attributed to the inherent demands placed on singers at faster tempos. The necessity for rapid transitions from note to note during high-BPM voices likely reduces the duration available for pitch fluctuation within individual notes, thereby contributing to enhanced pitch stability. Conversely, this accelerated pace simultaneously increases the cognitive and motor load required for accurate pitch control. Such an elevated processing demand makes precise adjustment to the target pitch more challenging, which can result in the overall deviation from the reference pitch becoming difficult to correct, or potentially even worsening.

C. Lyrics

The effect of lyrical differences was small when contrasted with other factors. For $MedAE$, in conditions ranging from 3-1-1 to 3-1-3 and 3-2-1 to 3-2-3, Lyrics 2 exhibited a smaller median value. Conversely, for $StdDev_{cent}$, almost no significant difference was observed between Lyrics 1 and 2.

While such differences were observed, their impact was limited compared to those caused by melodic leaps and varying BPMs. Therefore, it is considered that within the scope of the relatively simple syllable structures used in this experiment for fundamental frequency analysis, articulatory movement itself did not become a primary difficulty factor for pitch control.

VII. Conclusion

In this study, we investigated the influence of melodic structure (degree of leap), tempo (BPM), and lyrics on pitch stability (low fluctuation of pitch within a note) and accuracy (closeness of pitch to a reference voice) during singing, using data from 100 singers.

The results yielded the following insights:

- The degree of leap is the most influential factor on singing voice; as the leap becomes larger, both pitch stability and accuracy significantly decrease. This indicates that large leaps impose a high cognitive and motor load on the singer.
- The influence of BPM showed different patterns for stability and accuracy. While the first quartile, median, and third quartile for pitch stability within a note tended to decrease across all conditions from BPM 80 to 150, those for overall pitch accuracy increased monotonically. This suggests that participants who were not singing accurately struggled even more. Furthermore, the decrease in the third or median quartile between BPM 120 and BPM 150 suggests that familiarity could be a factor.
- The influence of lyric type was small compared to the other two factors above. Therefore, within the scope of the relatively simple syllable structures used in this experiment, it is thought that articulatory movement itself did not become a primary source of difficulty for pitch control.

In the future, we will conduct a more detailed analysis of leaps and BPM, which were found to have a particularly significant impact on singing voice in this experiment. Additionally, we will further investigate the influence of familiarity, which was not fully captured this time, and the effects of more complex rhythmic patterns on singing accuracy.

Furthermore, we will clarify how individual differences, such as the singer's musical experience and skill level, interact with the effects of these factors (leaps, BPM, rhythm, etc.). Moreover, through verification using songs with more diverse musical characteristics than those used in this study, we aim to enhance the generalizability of our findings.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Numbers 22H03711 and 24H00748.

References

[1] Y. Yamamoto and Y. Hiraga, "A study on calculation of singing difficulty of popular songs for a song search system focusing on singability and difficulty of singing," (in Japanese), in *The Special Interest Group Technical Reports of Information Processing Society of Japan*, vol. 2019-MUS-124, 2019, pp. 1–6.

[2] L. Ou, X. Ma, M.-Y. Kan, and Y. Wang, "Singable and controllable neural lyric translation," in *arXiv*, vol. 2305.16816, 2023.
[3] H. Kim, K. Watanabe, M. Goto, and J. Nam, "A computational evaluation framework for singable lyric translation," in *arXiv*, vol. 2308.13715, 2024.
[4] S. Chen, "A brief analysis of singability in erhu fiddle performance and teaching," in *Curriculum and Teaching Methodology*, vol. 5, 2022.
[5] M. D. Barone, K. M. Ibrahim, C. Gupta, and Y. Wang, "Empirically weighting the importance of decision factors for singing preference," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 529–536.
[6] V. Sébastien, H. Ralambondrainy, O. Sébastien, and N. Conruyt, "Automatically determining scores difficulty level for instrumental e-learning," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 571–576.
[7] Q. Liang, X. Ma, F. Doshi-Velez, B. Lim, and Y. Wang, "Improving the singability of AI-generated lyrics with prosody explanations," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 7877–7885.
[8] L. Ou, X. Ma, and Y. Wang, "Joint learning of wording and formatting for singable melody-to-lyric generation," in *arXiv*, vol. 2307.02146, 2024.