

## A APPENDIX

This supplementary material describes the dataset examples, experiment setups, and character length distribution.

### A.1 DATASET EXAMPLE

Here, we present the several English samples we collected using GPT-4 requests in Table 6. The "Possible Pronunciation" is necessary for the TTS models to generate speech and is extremely helpful for the human speech annotators as they can use it for reference if they do not know how to read the equation properly and simplifies the criteria for the human annotator.

### A.2 ADDITIONAL METRICS

Let us recall the main metrics in more detail.

Character Error Rate (CER) which is defined as the ratio of the normalized edit distance (Levenshtein distance) between the predicted sequence and the ground truth, normalized by the total number of characters in the reference:  $CER = \frac{S+D+I}{N}$ , where  $S$  is the number of substitutions,  $D$  is the number of deletions,  $I$  is the number of insertions, and  $N$  is the total number of characters in the reference.

The Word Error Rate (WER) is defined similarly to the CER but considers words instead of characters. CER and WER are commonly used in ASR tasks.

ROUGE-1 calculates the unigram recall between the predicted output and the reference text.

$$ROUGE-1 = \frac{\sum_{\text{unigram} \in \text{ref}} \min(\text{count}(\text{unigram}), \text{count}(\text{unigram\_pred}))}{\sum_{\text{unigram} \in \text{ref}} \text{count}(\text{unigram})} \quad (1)$$

This metric is widely used for summarization and transcription tasks to evaluate the lexical overlap between predicted and reference outputs.

BLEU and sacreBLEU evaluate n-gram precision by comparing the predicted output against the reference. BLEU is computed as:

$$BLEU = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (2)$$

where BP is the brevity penalty,  $p_n$  is the precision of n-grams, and  $w_n$  are weights. SacreBLEU applies different tokenization (Papineni et al., 2002; Post, 2018).

chrF and chrF++ are character-based F-scores metrics that compute a balance between precision and recall at the character level.

We’ve tried to train LLM with pronunciations from all 5 ASR systems from Table 1 to make it an ASR-agnostic model, but the model’s accuracy was worth more than just with Whisper. For results see Tables 7 and 8

As we can see, Qwen’s performance has declined, so mixing Audio models in this way is a bad idea because the training examples could be inconsistent.

Let’s measure case-sensitive performance when  $\phi$  and  $\Phi$  mean different symbols. See Tables 9 and 10

As we can see, the performance drop of the models was not as severe. This means that models in general, and Salamon in particular, were trained well, and data in terms of capitalized and non-capitalized symbols was well labelled.

The rest of the lower-cased metrics. This will be an edition to Table 2. For more details see Table 11

In this case, we also trained some models on non-overlapping formulas to check performance boost or degradation. See Tables 3 and 12.

And the second table for cross-language performance boost (all discussion was made 5.1). For the remaining metrics, see Table 13.

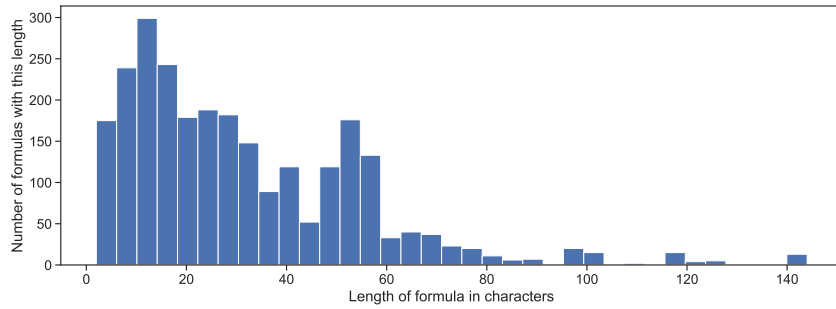


Figure 4: Distribution of equations lengths in the S2L dataset.

### A.3 CHARACTER LENGTH DISTRIBUTION

We provide additional information about our dataset. Mainly, it is num of characters in the formula in our test set for English. See Figure 4.

Here, we can see that our models were trained to predict both short and long equations with up to 140 characters.

Table 6: Example of the dataset samples for further annotation by speaker and TTS models.

Topic	Possible Pronunciation	Equation
Numbers	Fraction: 2 over 5	$\frac{2}{5}$
Calculus. Integrals	Integral: integral of x cubed dx equals x to the fourth over 4 plus constant	$\int x^3 dx = \frac{x^4}{4} + C$
Basic Geometry	the distance between two points (x1, y1) and (x2, y2) is the square root of (x2 minus x1) squared plus (y2 minus y1) squared	$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
Basic Functions	f of x is equal to x minus 3 divided by x squared minus 9	$f(x) = \frac{x-3}{x^2-9}$
Partial Derivatives	The partial derivative of f with respect to x and then y equals d squared f divided by d x d y	$\frac{\partial^2 f}{\partial x \partial y}$
Linear Algebra	the cross product of vectors a and b is a vector perpendicular to both	$a \times b$
Statistics	phi of x is equal to one divided by the square root of two pi times e to the power minus x squared divided by two	$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Complex Analysis	the exponential function of z is e raised to the power z	$\exp(z)$
Differential Equations	the solution to d y over d x equals negative k y is y equals c e to the negative k x	$\frac{dy}{dx} = -ky$ is $y = Ce^{-kx}$
Field Theory	the electromagnetic field tensor is given by F mu nu equals partial mu A nu minus partial nu A mu	$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$
Quantum Mechanics	the Schrödinger equation for a free particle is i h bar d psi over d t equals minus h bar squared over 2 m d squared psi over d x squared	$i\hbar \frac{d\psi}{dt} = -\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2}$
QFT	the Lagrangian density for the gauge field is minus one over four F mu nu F mu nu	$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$
Particle Physics	the mass of the Z boson is approximately 91.2 GeV/c squared	$m_Z \approx 91.2 \text{ GeV}/c^2$
Machine Learning	the F1 score with precision 0.75 and recall 0.5	$F1 = 2 \frac{0.75 \times 0.5}{0.75 + 0.5} = 0.6$
General Physics	Period of a pendulum: two pi times square root of length divided by gravitational acceleration	$T = 2\pi \sqrt{\frac{L}{g}}$
Trigonometry	Euler’s formula, e to the power i times pi plus one equals zero	$e^{i\pi} + 1 = 0$
Thermodynamics	Gibbs free energy, G equals H minus TS	$G = H - TS$

Table 7: Metrics results (%) for Qwen trained with 5 ASR models

Model	CER ↓	Rouge-1 ↑	sBLEU ↑	chrF ↑
Qwen2.5-0.5B	43.21	78.49	50.06	60.35

Table 8: Remaining metrics results (%) for Qwen trained with 5 ASR models

Model	WER ↓	METEOR ↑	BLEU ↑	chrF++ ↑
Qwen2.5-0.5B	75.33	57.21	47.06	58.88

Table 9: Case-sensitive metrics (%) for different Language Models

Model	Language	CER ↓	Rouge-1 ↑	sBLEU ↑	chrF ↑
Qwen2.5-0.5B	Eng	45.79	77.78	50.46	61.06
Qwen2.5-Math-1.5B	Eng	<b>44.39</b>	79.29	51.02	61.67
SALMONN-13B	Eng	44.47	<b>83.88</b>	<b>56.76</b>	<b>66.70</b>
Flan-T5	Eng	67.52	53.47	10.43	26.01
Qwen-Audio	Eng	54.64	76.63	54.79	57.61
Qwen2.5-0.5B	Rus	13.45	89.71	72.67	85.47
SALMONN-13B	Rus	<b>10.59</b>	<b>93.59</b>	<b>76.52</b>	<b>91.38</b>
Qwen2.5-0.5B	Eng+Rus	<b>23.39</b>	86.22	66.26	78.74
SALMONN-13B	Eng+Rus	24.99	<b>89.93</b>	<b>68.69</b>	<b>82.82</b>

**Note:** All formulas in the training and validation sets were voiced using TTS, and those in the test set were voiced with real speakers.

Table 10: Remaining case-sensitive metrics (%) for different Language Models

Model	Language	WER ↓	METEOR ↑	BLEU ↑	chrF++ ↑
Qwen2.5-0.5B	Eng	79.60	56.89	47.16	59.44
Qwen2.5-Math-1.5B	Eng	76.78	57.52	47.85	60.24
SALMONN-13B	Eng	<b>72.20</b>	<b>61.91</b>	<b>53.08</b>	<b>65.06</b>
Flan-T5	Eng	111.83	20.47	6.19	24.84
Qwen-Audio	Eng	102.91	53.67	42.53	55.89
Qwen2.5-0.5B	Rus	28.14	80.78	70.55	83.68
SALMONN-13B	Rus	<b>18.13</b>	<b>84.91</b>	<b>74.95</b>	<b>90.09</b>
Qwen2.5-0.5B	Eng+Rus	42.46	73.63	63.80	78.18
SALMONN-13B	Eng+Rus	<b>40.02</b>	<b>77.24</b>	<b>66.77</b>	<b>81.38</b>

**Note:** All formulas in the training and validation sets were voiced using TTS, and those in the test set were voiced with real speakers.

Table 11: Remaining lower-case metrics (%) for different Language Models

Model	Language	WER ↓	METEOR ↑	BLEU ↑	chrF++ ↑
Qwen2.5-0.5B	Eng	76.85	56.89	50.42	62.71
Qwen2.5-Math-1.5B	Eng	69.16	60.33	55.57	66.77
ProofGPT-1.3B	Eng	69.64	55.86	49.73	62.50
SALMONN-13B	Eng	<b>68.90</b>	<b>61.91</b>	<b>57.55</b>	<b>69.20</b>
InternLM2-1.8B	Eng	81.01	57.30	50.65	62.55
Flan-T5	Eng	109.26	20.47	7.69	27.53
Qwen-Audio	Eng	100.18	53.67	45.67	59.10
Qwen2.5-0.5B	Rus	27.14	80.78	70.64	84.34
Qwen2.5-Math-1.5B	Rus	23.80	81.65	72.03	86.47
ProofGPT-1.3B	Rus	32.14	79.10	68.51	82.22
SALMONN-13B	Rus	<b>17.94</b>	<b>84.91</b>	<b>75.05</b>	<b>90.36</b>
Qwen2.5-0.5B	Eng+Rus	41.47	73.63	64.75	78.18
ProofGPT-1.3B	Eng+Rus	43.26	72.20	62.94	76.37
SALMONN-13B	Eng+Rus	<b>38.80</b>	<b>77.24</b>	<b>67.85</b>	<b>82.62</b>

Table 12: Remaining metrics (%) results on overlapping formulas on train, val and test sets

Model	Language	Test	WER ↓	METEOR ↑	BLEU ↑	chrF++ ↑
Qwen2-0.5B	Rus	Human	14.82	86.74	78.46	91.87
Qwen2.5-0.5B	Rus	Human	<b>13.91</b>	<b>86.77</b>	<b>78.77</b>	<b>91.92</b>
Qwen2-0.5B	Eng	Human	40.37	73.88	68.60	76.53
Qwen2.5-0.5B	Eng	Human	<b>38.54</b>	<b>74.59</b>	<b>69.71</b>	76.53
Qwen2-0.5B	Eng+Rus	Human	<b>57.02</b>	<b>68.83</b>	<b>58.82</b>	70.78
Qwen2.5-0.5B	Eng+Rus	Human	58.27	68.56	58.60	<b>70.85</b>

**Note:** All formulas in the training and validation sets were voiced using TTS, and those in the test set were voiced with real speakers.

Table 13: Remaining metrics (%) results on non-overlapping cross-language fine-tuning

Model	Train Language	Test Language	WER ↓	METEOR ↑	BLEU ↑	chrF++ ↑
Qwen2-0.5B	Eng	Eng	<b>40.37</b>	<b>73.88</b>	68.60	76.53
Qwen2-0.5B	Eng+Rus	Eng	40.38	75.86	<b>70.71</b>	<b>78.59</b>