Appendix

This supplementary material describes the dataset examples and presents additional results.

Dataset Example

An overview of the dataset collection pipeline is presented in Figure 2.

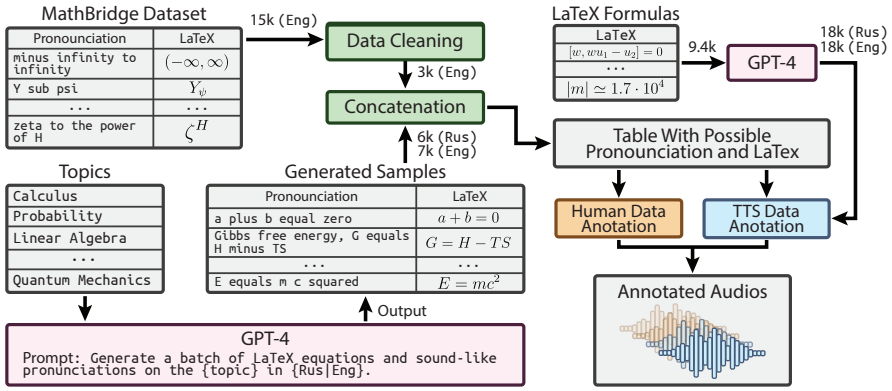

Figure 2: `S2L-equations` collection and annotation pipeline overview.

Let us compare transcriptions of 5 ASR models for one particular human-annotated audio in Table 8.

Table 8: Example of transcription of one particular human-annotated audio of the $\nabla_\nu A^\mu = \frac{\partial A^\mu}{\partial x^\nu} + \Gamma^\mu_{\nu\rho} A^\rho$ equation.

Model	Transcription
Whisper-L	The covariant derivative of a vector a mu equals partial mu with respect to X nu plus gamma upper rho mu nu times a rho.
WavLM	the covarient derivative of a vector a mou equals partial moo with respect to ex new plus gama upper row moo new times a row
Wav2Vec2	the covariant derivative of a vector a mu equals partial mo with respect to x-new plus scamma upper row mo new times a row
Qwen-audio	The covariant derivative of a vector mu equals partial mu with respect to x nu plus gamma upper row mu nu times a row
Canary	The covariant derivative of a vector amu equals partial amu with respect to x nu plus gamma upper rho moon nu times a rho.

Let us present several English samples we collected using GPT-4 requests in Table 9. The "Posssible Pronounciation" 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.

For the `S2L-sentences`, let us illustrate the evaluation challenge. Consider the CER between the predicted sequence "Given a fixed graph F, a typical problem on a large graph G on n vertices that contains no copy of F can have an upper bound on the number of its edges, denoted by $X(n, F)$" and the ground-truth "Given a fixed graph F, a typical problem in extremal graph theory asks for the maximum number of edges that a large graph G on n vertices containing no copy of F can have, denoted by ex(n, F)." The equation-only CER is 27.27%.

Metrics Description

We proceed by examining the primary and additional metrics in detail.

Table 9: Example of the dataset samples for further annotation by speaker and TTS models.

Topic	Possible Pronunciation	Equation
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	$a \times b$
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	$m_Z \approx 91.2\,\text{GeV}/c^2$
General Physics	Period of a pendulum: two pi times square root of length divided by gravitational acceleration	$T = 2\pi\sqrt{\frac{L}{g}}$
Mathematical Physics	Bessel function of the first kind of order zero, j sub zero is equal to the sum from m equals zero to infinity, of minus one to the power m, divided by m factorial squared, times x divided by two to the power of 2 m	$J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2}\left(\frac{x}{2}\right)^{2m}$
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	$G = H - TS$

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:

$$\text{CER} = \frac{S + D + I}{N}, \tag{1}$$

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.

$$\text{ROUGE-1} = \frac{\sum_{\text{unigram} \in \text{ref}} \min(\text{count}(\text{unigram}), \text{count}(\text{unigram\_pred}))}{\sum_{\text{unigram} \in \text{ref}} \text{count}(\text{unigram})} \tag{2}$$

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:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

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). TeXBLEU is a variant of the BLEU score adapted to evaluate LaTeX string generation tasks, particularly mathematical expressions. It penalizes syntactic mistakes and helps measure the quality of generated LaTeX code.

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

$$\text{chrF}_\beta = (1 + \beta^2) \cdot \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}, \tag{4}$$

Where chrP and chrR represent the arithmetic mean of character n-gram precision and recall across all n-grams; chrP is the percentage of character n-grams in the hypothesis that also appear in the reference, and chrR is the percentage of character n-grams in the reference that are also found in the hypothesis. chrF++ is chrF for $n = 2$.

TeXBleu is relatively insensitive to the significant errors.

Training Hyperparameteres

The default loss function was cross-entropy, and the default optimizer was AdamW (Loshchilov, Hutter et al. 2017). Qwen models for `S2L-Equations` experiments were trained on 1 A100 GPUs for 1 epoch, and batch size was set to 16 samples per batch. AdamW optimizer was used with weight decay of 0.01 with a learning rate $1e-4$ and linear learning rate scheduler.

For `S2L-Sentences` experiments, Qwen models were trained on a single A100 GPU for 1 epoch, and the batch size was set to 16 samples per batch. AdamW optimizer was used with weight decay of 0.01 with learning rate $1e-4$ and linear learning rate scheduler.

SALMONN was trained with the LoRA technique on Llama. Target modules were set to attention layers, rank was 8, alpha parameter was 32, and dropout was set to 10%. Whisper and Beats models were frozen. The model was trained on Nvidia H100-80Gb 2 GPUs for 6 epochs. The learning rate was set to 3e-5 with a warm-up for 3000 steps and cosine decay. Gradient accumulation was set to 3 iterations. The batch size was set to 12 samples per batch. Automated mixed precision with float16 was used.

For the few-shot learning, the following system prompt and examples were used:

```
1
2  system_prompt = """You are a mathematics text processing assistant. Your task is to
       convert informal mathematical expressions into proper LaTeX format while keeping all
       other text unchanged. Maintain the original structure and wording, only modifying
       mathematical notation. Respond only with the processed text, no additional commentary.
       """
3
4  % few_shot_examples = [
5      {'user': 'Our first result is an explicit description of F sub Lin in the case where m
           equals k square bracket delta is the Stanley-Reisner ring of a simplicial complex
           delta.',
6       'assistant': 'Our first result is an explicit description of $\\Fblin$ in the case
           where $M = \\Bbbk[\\Delta]$ is the Stanley-Reisner ring of a simplicial complex $
           \\Delta$.'},
7
8      {'user': 'The results show the cystinatic improvability with increasing N sub Ln
          raised to G.',
9       'assistant': 'The results show the systematic improvability with increasing $n_\\
           mathrm{mom}^G$.'},
10
11     {'user': 'or any D divides and denote by or superscript D of L the set of orientations
           with isotropy group isomorphic to the cyclic group of order D i.e.',
12      'assistant': 'For any $d \\mid m$ denote by $\\Or^{d}(L)$ the set of orientations
           with isotropy group isomorphic to $\\mathbb{Z}_d$, i.e.'},
```

```
13
14      {'user': 'Hence, see what is a subscript of and minus two wins by the tie-breaking.',
15       'assistant': 'Hence, $c_{n-2}$ wins by the tie breaking.'},
16
17      {'user': 'Let A from Y to X be amorphous between smooth varieties over field K.',
18       'assistant': 'Let $f: Y\\longrightarrow X$ be a morphism between smooth varieties
            over $k$.'}
19  ]
```

Additional Results

Additional Results for the Main Text Tables

Full version of the Table 4 from the main text for `S2L-equations` results is Table 10. Additional few-show results for the `S2L-sentences` (Table 7) are presented in Table 11.

Additional Models

We also evaluated InternLM, ProofGPT, and FlanT5 on the subsets of `S2L-equations` on additional experiments with different splits. Results are presented in Table 12. ProofGPT-1.3B demonstrated good performance, except for the Russian language. In the setting when train and test data both have a mix of genuine and artificial audio, and the test set equations have no overlapping with the equations from the train, SALMONN-13B demonstrates the best metrics except CER on all languages, while Qwen2.5 has a slight edge over SALMONN regarding CER. For instance, on the English subset, SALMONN leads with the highest Rouge-1 (83.88), sBLEU (60.68), and chrF (71.04) scores. However, its CER (42.42) is slightly higher than Qwen2.5-Math-1.5B, which has the lowest CER (39.54) and ranks second in Rouge-1 (81.43) and chrF (68.34). The Qwen-Audio performs worse than other methods, probably due to re-implementation nuances. The second part of the table compares Qwen2-0.5B and Qwen2.5-0.5B for English and Russian languages for the random and disjoint (test equations do not overlap train ones) splits. For both languages, Qwen2.5-0.5B consistently outperforms Qwen2-0.5B in terms of Rouge-1 and sBLEU. Interestingly, in the case of the combined English and Russian datasets, the 2 models exhibit very close performance, with Qwen2.5-0.5B showing marginal improvements in accuracy metrics while having a slightly higher CER.

Rest of the metrics

We tried to train LLM with pronunciations from all 5 ASR systems from Table 8 to make it an ASR-agnostic model, but the model's accuracy was worth more than just with Whisper. For results see Table 13.

We measured case-sensitive performance (for example, ϕ and Φ mean different symbols). Results are presented in Tables 14 and 15. As we can see, the performance drop is not as severe. This generally means that models were trained well and that data regarding capitalized and non-capitalized symbols was labelled well. The rest of the metrics from the Table 12 are represented in Table 16 as an addition with the lower-cased metrics for the `S2L-eqautions` part.

Cross-Language Learning.

One of the advantages of fine-tuning multilingual language models is the ability to extract information from one language that is not available in another. For example, LaTeX special symbols `\simeq` and `\hat` are not presented in the Russian part of the equations dataset but in English. Qwen2.5, trained in English and Russian, can transcribe "approximately equal" in Russian to `\simeq` ($\simeq$). Another observation is that the models are primarily English-oriented, so Qwen2.5-Math-1.5B and Qwen2-0.5B trained in Russian can generate only simple formulas in English. The reverse situation works worse - Qwen2.5-0.5B, trained in English, cannot perform post-correction in Russian.

Table 10: `S2L-equations` results. Disjoint split: test equations do not overlap with train equations. "A": artificially (TTS) annotated audio except 400k samples extracted from `MathBridge`; "H": human-annotated audio; "Mix": combination of "A" and "H"; CER is calculated for lower-case. "Q-αB" and "Q-math-αB" stand for Qwen2.5-αB-instruct and Qwen2.5-math-αB-instruct, respectively. "Full" implies addition of 400k artificially annotated samples from `MathBridge` to the "A" set.

Model	Train	Train Language	Test Language	Test: Mix		Test: H		Test: A	
				CER	TeXBLEU	CER	TeXBLEU	CER	TeXBLEU
MathSpeech	`MS-train`	Eng	Eng	64.04	83.71	59.32	83.64	69.65	83.80
Q-0.5B	A	Eng	Eng	33.28	88.61	33.26	88.54	33.30	88.70
Q-0.5B	A	Eng+Rus	Eng	34.78	87.90	34.94	87.57	34.59	88.31
Q-0.5B	H	Eng	Eng	36.91	87.86	35.01	88.25	39.16	87.38
Q-0.5B	H	Eng+Rus	Eng	35.43	88.06	33.94	88.47	37.19	87.56
Q-0.5B	Mix	Eng	Eng	31.41	88.83	31.06	88.87	31.82	88.78
Q-0.5B	Mix	Eng+Rus	Eng	32.33	88.60	31.18	88.89	33.69	88.24
Q-0.5B	Mix-full	Eng+Rus	Eng	27.21	90.20	27.03	90.14	27.42	90.27
Q-1.5B	A	Eng	Eng	31.24	89.22	31.37	89.15	31.08	89.31
Q-1.5B	A	Eng+Rus	Eng	30.73	88.92	30.70	88.73	30.77	89.16
Q-1.5B	H	Eng	Eng	29.69	89.41	27.57	89.69	32.22	89.07
Q-1.5B	H	Eng+Rus	Eng	30.93	89.04	28.85	89.42	33.39	88.57
Q-1.5B	Mix	Eng	Eng	29.76	89.28	28.93	89.44	30.74	89.09
Q-1.5B	Mix	Eng+Rus	Eng	31.14	89.37	30.08	89.43	32.40	89.28
Q-1.5B	Mix-full	Eng+Rus	Eng	25.69	90.70	24.91	90.74	26.61	90.66
Q-math-1.5B	A	Eng	Eng	29.44	89.61	30.00	89.33	28.77	89.96
Q-math-1.5B	A	Eng+Rus	Eng	29.57	90.00	29.44	89.80	29.74	90.23
Q-math-1.5B	H	Eng	Eng	30.16	89.83	28.97	90.13	31.58	89.46
Q-math-1.5B	H	Eng+Rus	Eng	31.45	89.25	30.71	89.43	32.34	89.02
Q-math-1.5B	Mix	Eng	Eng	28.53	89.97	28.08	90.13	29.05	89.76
Q-math-1.5B	Mix	Eng+Rus	Eng	27.75	89.85	27.54	89.89	28.01	89.79
Q-math-1.5B	Mix-full	Eng+Rus	Eng	25.01	90.90	25.05	90.90	24.97	90.89
Q-7B	A	Eng	Eng	28.15	90.10	28.07	89.96	28.25	90.26
Q-7B	A	Eng+Rus	Eng	27.32	90.12	26.16	90.20	28.70	90.03
Q-7B	H	Eng	Eng	27.97	89.99	26.93	90.29	29.20	89.62
Q-7B	H	Eng+Rus	Eng	26.89	90.18	26.43	90.36	27.44	89.95
Q-7B	Mix	Eng	Eng	26.10	90.58	25.80	90.68	26.46	90.45
Q-7B	Mix	Eng+Rus	Eng	27.78	90.11	26.55	90.37	29.24	89.78
Q-7B	Mix-full	Eng+Rus	Eng	26.17	90.50	25.96	90.51	26.43	90.48
Qwen-Audio	Mix	Eng	Eng	71.67	83.55	104.19	78.45	33.06	89.84
SALMONN	Mix-full	Eng	Eng	17.50	93.68	18.17	93.64	16.70	93.72
Gemma 3n	Mix-full	Eng	Eng	34.24	89.15	33.24	89.23	35.42	89.06
Q-0.5B	A	Rus	Rus	32.06	91.73	40.95	94.13	27.62	90.53
Q-0.5B	A	Eng+Rus	Rus	9.63	95.14	17.78	96.34	5.56	94.54
Q-0.5B	H	Rus	Rus	15.24	96.97	15.87	96.76	14.92	97.08
Q-0.5B	H	Eng+Rus	Rus	6.77	97.74	13.97	97.20	3.17	98.01
Q-0.5B	Mix	Rus	Rus	15.34	98.50	15.24	97.21	15.40	99.14
Q-0.5B	Mix	Eng+Rus	Rus	13.02	97.32	14.60	96.78	12.22	97.59
Q-0.5B	Mix-full	Eng+Rus	Rus	8.15	97.03	15.56	96.68	4.44	97.21
Q-1.5B	A	Rus	Rus	10.05	96.77	19.37	96.60	5.40	96.85
Q-1.5B	A	Eng+Rus	Rus	6.14	96.62	14.60	96.74	1.90	96.56
Q-1.5B	H	Rus	Rus	4.66	99.38	8.25	99.49	2.86	99.33
Q-1.5B	H	Eng+Rus	Rus	14.50	97.38	14.60	96.67	14.44	97.73
Q-1.5B	Mix	Rus	Rus	14.60	95.90	10.79	98.00	16.51	94.85
Q-1.5B	Mix	Eng+Rus	Rus	4.02	99.20	11.75	97.68	0.16	99.96
Q-1.5B	Mix-full	Eng+Rus	Rus	4.55	98.89	13.33	96.74	0.16	99.96
Q-math-1.5B	A	Rus	Rus	6.03	98.94	17.14	97.02	0.48	99.89
Q-math-1.5B	A	Eng+Rus	Rus	11.01	98.04	13.33	97.08	9.84	98.52
Q-math-1.5B	H	Rus	Rus	13.23	96.51	2.86	99.36	18.41	95.08
Q-math-1.5B	H	Eng+Rus	Rus	12.49	97.45	11.75	97.78	12.86	97.28
Q-math-1.5B	Mix	Eng+Rus	Rus	5.50	98.90	13.65	97.39	1.43	99.66
Q-math-1.5B	Mix	Rus	Rus	17.25	97.24	12.70	97.70	19.52	97.01
Q-math-1.5B	Mix-full	Eng+Rus	Rus	13.33	97.86	14.29	96.72	12.86	98.43
Q-7B	A	Rus	Rus	3.70	99.19	10.79	97.66	0.16	99.96
Q-7B	A	Eng+Rus	Rus	6.14	98.39	16.83	96.46	0.79	99.35
Q-7B	H	Rus	Rus	5.40	99.29	6.98	99.33	4.60	99.27
Q-7B	H	Eng+Rus	Rus	21.59	96.73	14.92	97.00	24.92	96.60
Q-7B	Mix	Rus	Rus	1.59	99.57	4.44	98.78	0.16	99.96
Q-7B	Mix	Eng+Rus	Rus	4.66	99.09	13.65	97.36	0.16	99.96
Q-7B	Mix-full	Eng+Rus	Rus	7.94	97.74	14.92	96.87	4.44	98.17
Flamingo 3	Mix	Rus	Rus	2.01	99.88	0.00	100.00	3.02	99.82
SALMONN	Mix-full	Rus	Rus	9.38	97.73	6.51	99.55	10.81	96.82
Gemma 3n	Mix-full	Rus	Rus	15.30	97.70	11.26	98.17	17.33	94.48

Table 11: `S2L-sentences` results for Few-Shot experiments. Disjoint split: test sentences do not overlap with train sentences. "A": artificially (TTS) annotated audio; "H": human-annotated audio; "Mix": combination of "A" and "H". CER is calculated for lower-case. "Q-αB" and "Q-math-αB" stand for Qwen2.5-αB-instruct and Qwen2.5-math-αB-instruct, respectively. "Sent." stands for sentence: metric calculated over the whole sentence; "Eq": only for the embedded equations; "Text": only for the text parts of the sentence.

Model	Train	Test: H				Test: A			
		CER			TeXBLEU	CER			TeXBLEU
		Sent.	Text	Eq.	Eq.	Sent.	Text	Eq.	Eq.
5-shot									
Q-0.5B	A	35.80	35.76	66.22	66.87	34.21	34.20	68.30	60.41
Q-0.5B	H	32.09	29.69	71.95	65.31	31.83	30.16	73.71	58.47
Q-1.5B	A	26.16	20.62	58.78	76.84	26.64	22.26	60.47	70.64
Q-1.5B	H	27.94	20.50	63.99	76.78	28.70	23.69	64.74	70.83
Q-math-1.5B	A	35.94	35.99	62.73	71.66	41.90	43.85	65.75	64.53
Q-math-1.5B	H	42.52	42.16	73.36	68.23	44.63	45.15	78.05	61.07
Q-7B	A	23.83	18.44	56.25	77.88	23.63	19.57	56.03	72.64
Q-7B	H	24.19	17.56	57.91	77.97	23.44	18.76	55.88	73.31
25-shot									
Q-0.5B	A	28.74	25.97	61.87	73.77	28.15	25.89	65.14	66.53
Q-0.5B	H	30.67	27.30	63.44	73.93	31.47	27.37	72.45	66.36
Q-1.5B	A	23.65	17.55	56.84	78.42	24.10	18.82	58.77	72.39
Q-1.5B	H	24.05	17.26	56.77	78.57	24.49	18.93	57.61	73.43
Q-math-1.5B	A	37.65	29.11	88.22	75.14	36.74	27.56	95.89	69.09
Q-math-1.5B	H	30.58	24.43	67.93	75.67	31.26	27.51	67.43	68.44
Q-7B	A	21.22	15.85	50.43	79.88	21.65	16.97	51.97	74.65
Q-7B	H	20.00	14.23	47.12	80.64	20.73	16.38	48.21	75.14

Table 12: `S2L-equations` (subset) results. SALMONN represent end-to-end Audio-LLMs, while all other models use ASR post-correction via fine-tuned LLMs. "A" denotes artificially (TTS) annotated audio, "H" refers to human-annotated audio, and "Mix" indicates a combination of both. "Rand" indicates a random split where equation-pronunciation-speaker/voice triplets are non-overlapping across train, validation, and test sets. "Disj" specifies a disjoint split where test equations do not appear in the training set.

Model	**Lang**	**Train**	**Test**	**Split**	**CER↓**	**Rouge-1↑**	**sBLEU↑**	**chrF↑**
Qwen2.5-0.5B	Eng	Mix	Mix	Disj	43.87	77.78	53.33	64.48
Qwen2.5-Math-1.5B	Eng	Mix	Mix	Disj	**39.54**	81.43	57.86	68.34
ProofGPT-1.3B	Eng	Mix	Mix	Disj	41.60	78.04	52.31	64.30
InternLM2-1.8B	Eng	Mix	Mix	Disj	49.23	78.12	61.00	64.24
Flan-T5	Eng	Mix	Mix	Disj	64.92	53.47	11.98	28.78
SALMONN-13B	Eng	Mix	Mix	Disj	42.42	**83.88**	**60.68**	**71.04**
Qwen2.5-0.5B	Rus	Mix	Mix	Disj	13.19	89.71	72.78	86.09
Qwen2.5-Math-1.5B	Rus	Mix	Mix	Disj	10.49	90.66	74.25	88.11
ProofGPT-1.3B	Rus	Mix	Mix	Disj	16.48	87.82	70.82	84.04
SALMONN-13B	Rus	Mix	Mix	Disj	**10.45**	**93.59**	**76.63**	**91.63**
Qwen2.5-0.5B	Eng+Rus	Mix	Mix	Disj	**22.70**	86.22	67.14	79.87
ProofGPT-1.3B	Eng+Rus	Mix	Mix	Disj	23.93	84.85	65.33	78.18
SALMONN-13B	Eng+Rus	Mix	Mix	Disj	24.27	**89.93**	**69.62**	**84.10**
Qwen2-0.5B	Eng	A	H	Rand	25.05	86.56	70.39	76.91
Qwen2.5-0.5B	Eng	A	H	Rand	**23.56**	**86.92**	**71.37**	**77.88**
Qwen2-0.5B	Rus	A	H	Rand	**7.09**	94.44	79.59	**92.79**
Qwen2.5-0.5B	Rus	A	H	Rand	7.49	**94.58**	**79.88**	92.73
Qwen2-0.5B	Eng+Rus	A	H	Disj	**30.36**	83.52	61.72	72.20
Qwen2.5-0.5B	Eng+Rus	A	H	Disj	31.13	**83.60**	**61.73**	**72.22**

Table 13: S2L-equations (subset). Metrics results (%) for Qwen trained with 5 ASR models.

Model	CER↓	Rouge-1↑	sBLEU↑	chrF↑	WER↓	METEOR↑	BLEU↑	chrF++↑
Qwen2.5-0.5B	43.21	78.49	50.06	60.35	75.33	57.21	47.06	58.88

Table 14: S2L-equations (subset). Case-sensitive metrics (%) for different Language Models. "Mix" means a combination of human-annotated and TTS. Lang means the language of the train/validation/test splits.

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

Table 15: S2L-equations (subset). Remaining case-sensitive metrics (%) for different Language Models. "Mix" means combination of Human annotated and TTS. Lang means language of train/validation/test splits

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

Table 16: `S2L-equations` (subset). Remaining results of lower-case metrics (%) for different models. SALMONN represents the Multimodal approach, while the rest of the models represent ASR post-correction. "A" stands for artificially annotated audio (TTS), "H" – human annotated audio, "Mix" – the combination of both "A" and "H". "Disj" split means that test equations do not intersect with the train equations, and "Rand" split means that train-test split was made randomly over generated pairs and equations from train might occur in the test but should be pronounced with different speakers or TTS models.

Model	Lang	Train	Test	Split	WER↓	METEOR↑	BLEU↑	chrF++↑
Qwen2.5-0.5B	Eng	Mix	Mix	Disj	76.85	56.89	50.42	62.71
Qwen2.5-Math-1.5B	Eng	Mix	Mix	Disj	69.16	60.33	55.57	66.77
ProofGPT-1.3B	Eng	Mix	Mix	Disj	69.64	55.86	49.73	62.50
SALMONN-13B	Eng	Mix	Mix	Disj	**68.90**	**61.91**	**57.55**	**69.20**
InternLM2-1.8B	Eng	Mix	Mix	Disj	81.01	57.30	50.65	62.55
Flan-T5	Eng	Mix	Mix	Disj	109.26	20.47	7.69	27.53
Qwen2.5-0.5B	Rus	Mix	Mix	Disj	27.14	80.78	70.64	84.34
Qwen2.5-Math-1.5B	Rus	Mix	Mix	Disj	23.80	81.65	72.03	86.47
ProofGPT-1.3B	Rus	Mix	Mix	Disj	32.14	79.10	68.51	82.22
SALMONN-13B	Rus	Mix	Mix	Disj	**17.94**	**84.91**	**75.05**	**90.36**
Qwen2.5-0.5B	Eng+Rus	Mix	Mix	Disj	41.47	73.63	64.75	78.18
ProofGPT-1.3B	Eng+Rus	Mix	Mix	Disj	43.26	72.20	62.94	76.37
SALMONN -13B	Eng+Rus	Mix	Mix	Disj	**38.80**	**77.24**	**67.85**	**82.62**
Qwen2-0.5B	Rus	A	H	Rand	14.82	86.74	78.46	91.87
Qwen2.5-0.5B	Rus	A	H	Rand	**13.91**	**86.77**	**78.77**	**91.92**
Qwen2-0.5B	Eng	A	H	Rand	40.37	73.88	68.60	76.53
Qwen2.5-0.5B	Eng	A	H	Rand	**38.54**	**74.59**	**69.71**	76.53
Qwen2-0.5B	Eng+Rus	A	H	Disj	**57.02**	**68.83**	**58.82**	70.78
Qwen2.5-0.5B	Eng+Rus	A	H	Disj	58.27	68.56	58.60	**70.85**