Author's Accepted Manuscript. By permission. Equivalent in content, but not in format, to the final Version of Record in *Modelling and Simulation 2026 – Proc. European Simulation and Modelling Conference, ESM'2026*, published by EUROSIS – The European Multidisciplinary Society for Modelling and Simulation Technology. https://www.eurosis.org/cms/index.php?q=taxonomy/term/58#ESM
Full citation: J. K. Krajewski, A. Hinc, L. J. Chmielewski, and A. Orłowski. How a small fine-tuned model can rival a large general-purpose model in culinary tasks. In S. Hennequin, D. Roy, J. Petronijevic, P. Stief, and T. Zhang, editors, *Modelling and Simulation 2026 – Proc. European Simulation and Modelling Conference, ESM'2026*, pp. 101–105. Metz, France, 21-23 October, 2026. EUROSIS.
Last compiled on November 16, 2026.

HOW A SMALL FINE-TUNED MODEL CAN RIVAL A LARGE GENERAL-PURPOSE MODEL IN CULINARY TASKS

Jan K. Krajewski[1]
[1]Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

Leszek J. Chmielewski[1]
[1]Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland
leszek_chmielewski@sggw.edu.pl (corresponding author)

Aleksandra Hinc[1]
[1]Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

Arkadiusz Orłowski[1,2]
[1]Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland, and
[2]Institute of Mathematics and Cryptology, Military University of Technology – WAT, Warsaw, Poland

ABSTRACT

It was tested whether language models with at most one billion parameters can rival a large general-purpose model in a narrow, structured domain. Culinary tasks were used as a controlled benchmark covering recipe classification, nutritional-value prediction, instruction generation, and description generation. Three public small models, Qwen3-0.6B, GPT-2-Large, and Llama-3.2-1B, were compared with GPT-4.1 before and after LoRA fine-tuning on 30 000 recipes. Fine-tuned Llama-3.2-1B approached GPT-4.1 in generation and classification, and exceeded it in structured nutritional prediction. The results show that domain-specialised small models can be a practical alternative to frontier models in well-defined tasks.

KEYWORDS

Small language models, Fine-tuning, LoRA, Evaluation metrics, Culinary NLP

INTRODUCTION

Large language models owe much of their success to scale, but this scale also brings high computational, financial, and infrastructural costs. For many practical applications the central question is therefore not whether the largest available model performs well, but whether a much smaller model can become competitive after domain adaptation.

We examine this question in culinary tasks, a domain combining structured labels, numerical nutritional values, procedural text, and natural-language descriptions. The hypothesis is that a properly fine-tuned small language model ($\leq 1\,B$ parameters) can approach or exceed a large general-purpose model in such tasks. The study compares pretrained and LoRA-fine-tuned SLMs with GPT-4.1 under identical prompts and validation data. It is based on work reported in (Krajewski 2026).

STATE OF THE ART

Transformers (Vaswani et al. 2017) and subsequent pretrained models such as BERT (Devlin et al. 2019) and GPT-type systems (Yenduri et al. 2024) established large-scale pretraining as the dominant NLP paradigm. Meanwhile, the cost of training, fine-tuning, and deploying very large models has stimulated research into Small Language Models (SLMs) (Kim et al. 2025). We treat models up to one billion parameters as small.

Parameter-Efficient Fine-Tuning, especially Low-Rank Adaptation (LoRA) (Hu et al. 2022), makes such models easier to specialise, as only small adapter matrices are trained while the base model remains frozen. Previous studies (Łępicki and Kurek 2025) indicate that SLMs can be competitive in specialised tasks, but systematic sub-1B versus frontier-model comparisons across several structured tasks remain limited. Culinary NLP has also been explored for recipe-related tasks (Rita et al. 2024), (Sugioka et al. 2024), making it a suitable benchmark for testing scale against specialisation.

MODELS AND TASKS

The evaluated SLMs were Qwen3-0.6B (Yang et al. 2025), GPT-2-Large (Radford et al. 2019), and Llama-3.2-1B (Touvron et al. 2023). GPT-4.1 (OpenAI, Kumar et al. 2025) served as the large general-purpose reference model and was not fine-tuned.

Four tasks were considered: recipe classification, nutritional-value prediction for six nutrients, generation of cooking instructions from dish name and ingredients, and generation of short dish descriptions from instructions. Together they cover classification, regression-like numerical similarity, and free-form generation.

MEASURES OF QUALITY

The evaluation framework was designed for comparability across heterogeneous tasks: classification, structured numerical prediction, and free-form generation. Because

the tasks differ fundamentally in output structure, a unified metric is neither feasible nor desirable. Instead, each task category is evaluated with task-appropriate criteria under a shared experimental protocol.

Quantitative Measures and their Limitations

For multi-class classification, standard accuracy was used. Given the balanced distribution of categories in the validation set, accuracy provides a transparent and interpretable measure.

Text generation evaluation presents greater difficulty. Classical n-gram–based metrics such as BLEU, ROUGE, or METEOR penalize lexical and syntactic variation even when semantic equivalence is preserved (Hu and Zhou 2024). Culinary instructions frequently admit multiple equivalent formulations (e.g., reordering steps that are causally independent, using synonyms for ingredients or actions). In such cases, surface-level overlap underestimates true semantic similarity.

For this reason, semantic-level metrics were preferred. BERTScore (Zhang et al. 2020) was employed to compare contextual token embeddings between generated and reference texts. Additionally, cosine similarity between sentence embeddings (Yin and Zhang 2024) was computed to capture global semantic alignment. These metrics do not rely strictly on token overlap and better reflect functional equivalence in procedural instructions.

Two additional reference-free metrics were included. Perplexity (Chen et al. 2018) was measured to assess fluency and syntactic plausibility. Readability (Eleyan et al. 2020) was computed using established readability indices to detect excessive structural complexity. While neither directly measures culinary correctness, both provide useful constraints on linguistic quality.

Nutritional value prediction required a different approach. Exact numerical matching is inappropriate because minor deviations may be nutritionally negligible. Therefore, similarity was defined using a logarithmic-relative formulation:

$$p = e^{-|\ln((a+\varepsilon)/(b+\varepsilon))|} \tag{1}$$

where a denotes the predicted value, b the ground truth, and ε a stabilizing constant. The logarithmic term ensures scale invariance, penalizing proportional rather than absolute error. The exponential transformation bounds the similarity score within $(0, 1]$. Values above a threshold of 0.7 were considered sufficiently similar; this value was chosen to balance tolerance to minor deviations with practical nutritional relevance; pilot checks indicated stable model ranking under moderate threshold variations. The parameter ε was chosen empirically per nutrient ($\approx 10 - 15\%$ of the median nutrient value) to avoid instability for small magnitudes. This formulation provides three advantages: proportional fairness across nutrients with different scales, tolerance for small nutritional deviations, and interpretability of similarity as a continuous confidence measure.

Human-based Evaluation and its Potential Automation

Automatic metrics alone cannot guarantee culinary validity. A generated instruction may achieve high semantic similarity while containing procedural flaws (e.g., incorrect cooking order, omission of essential steps), so a structured human-evaluation protocol was introduced.

For instruction generation, six binary conditions were defined: linguistic correctness, culinary validity, absence of procedural errors, appropriate ingredient handling, logical completeness, and presence of time specification. Each condition was evaluated independently, enabling fine-grained analysis rather than a single aggregated score. The criteria are a result of observation of incorrect answers returned by the trained networks.

For description generation, analogous binary criteria were defined: semantic validity, linguistic correctness, proper length compression, lack of hallucinations, sensory informativeness, and inclusion of key dish features.

The protocol was constructed inductively by analyzing recurring failure modes in baseline outputs. Importantly, the criteria avoid subjective notions such as taste preference and instead focus on structural and procedural adequacy.

Although currently implemented as human annotation, the structure of the protocol suggests partial automatization potential. Several criteria (e.g., time specification, length constraints, lexical overlap with ingredients) can be programmatically detected. Others, such as logical consistency, may be approximated through rule-based or model-based validators. The protocol thus provides a foundation for future reference-free automatic culinary evaluation.

DATASETS USED

The experimental design relied on clearly separated data sets to prevent leakage and to ensure fair comparison between models.

Three distinct data sets were used: (i) pretrained corpora implicitly embedded in each model, (ii) a supervised fine-tuning dataset, and (iii) an independently constructed validation dataset.

THE TESTING ENVIRONMENT

Experiments were conducted in Python using PyTorch and Hugging Face Transformers (Transformers 2019). Fine-tuning was performed on a local system with RTX 4070 GPU, 12 GB VRAM. LoRA adapters were applied to attention projection matrices. Training times ranged from 1 to 10 hours depending on task.

Pretraining

All evaluated small models were used in their publicly available pretrained versions. Their original pretraining corpora are not identical and remain partially undisclosed in detail. Importantly, none of the models were originally specialized for structured culinary reasoning tasks such as nutrient prediction.
GPT-4.1 was evaluated in its publicly available inference configuration without domain-specific fine-tuning. This ensures that any performance gains observed in small models after specialization are attributable to controlled adaptation rather than hidden domain training.

Fine-tuning

Fine-tuning was performed using a dataset `recipe` (Shengtao 2019) of over 30 000 English-language recipes sourced from AllRecipes (Kappel et al. 2026). It contains fields including dish name, ingredient list, preparation instructions, nutritional values per serving, category labels, and metadata.
The dataset was selected for the following properties: balanced distribution across culinary categories, wide nutritional range (low- and high-calorie dishes), natural, non-synthetic language, minimal missing values, and consistent formatting. Records were transformed into instruction–target pairs tailored to each of the four defined tasks. Prompts were standardized to reduce variance induced by formatting differences. For generative tasks, instruction templates constrained output format (e.g., short answer requirements) to avoid uncontrolled verbosity. No record from this dataset was included in the validation set.
During the fine-tuning, the biggest challenges were the high risk of overfitting and the irregularity of the training data. The former was mitigated by using such LoRA techniques like relaxed warm-up and cool-down of the learning rate, as well as monitoring the loss function and recording several points of training progress. The latter was addressed by using a specific structure for each command and expected response, with specific tokens defining the start and stop of the target value.
By way of example of overtraining let us mention that one of the systems proposed to bake the scrambled eggs in an oven, like pancakes.

Validation and Experimental Control

A separate validation dataset with 150 manually curated recipes was constructed. Recipes were collected from multiple public culinary sources and manually verified. Each record contains: dish name, ingredient list with proportions, full preparation instructions, nutritional values per serving (calories, fat, cholesterol, carbohydrates, sugar, protein), assigned category label, and descriptive summary.

Table 1: Classification accuracy [%]

Metric \ Model	GPT-4.1	Qwen3-0.6B	GPT-2-Large	Llama-3.2-1B
		Basic / Fine-tuned		
1	2	3 / 4	5 / 6	7 / 8
Classification accuracy	68.5	52 / 54	37 / 65	11 / 67

Table 2: Instruction generation: automatic metrics

Metric \ Model	GPT-4.1	Qwen3-0.6B	GPT-2-Large	Llama-3.2-1B
		Basic / Fine-tuned		
1	2	3 / 4	5 / 6	7 / 8
BERTScore	0.45	0.26 / 0.26	0.25 / 0.24	0.22 / 0.34
Cosine embedding similarity	0.87	0.77 / 0.78	0.75 / 0.77	0.74 / 0.83
Perplexity	13.30	6.11 / 6.51	6.30 / 7.37	6.24 / 6.30
Readability	74.44	75.18 / 78.92	77.00 / 77.30	72.71 / 78.71

Nutritional values were computed proportionally using standardized USDA tables, for internal consistency.
The validation dataset was designed to maximize heterogeneity: multiple cuisines and cultural origins, varied preparation times, simple and multi-step dishes, both low- and high-calorie meals. This heterogeneity reduces the risk of biased evaluation toward specific recipe types. All models, including GPT-4.1, were evaluated on the identical validation dataset under the same prompt formulations. No gradient updates were performed on this dataset. This guarantees comparability and eliminates data leakage between training and evaluation phases.

RESULTS

This section presents experimental outcomes in three stages: (i) evaluation of the large reference model without domain adaptation, (ii) evaluation of pretrained small models, and (iii) evaluation after LoRA-based fine-tuning. All models were tested on the same validation dataset using identical prompt formulations.

Large Model – No Fine-tuning

GPT-4.1 was evaluated as a strong general-purpose baseline which establishes an upper reference bound for generative fluency and structured reasoning without domain-specific specialization. It achieved 68.5% accuracy on the curated validation dataset (Tab. 1, col. 2). Automatic metrics computed on generated instructions have been collected in Tab. 2, col. 2. Results of human evaluation of instruction generation have been shown in Tab. 3, col. 2. For description generation, the human-based evaluation yielded the percentages of outputs satisfying the assumed criteria listed in Tab. 4, col. 2.
Structured nutritional prediction revealed substantial weaknesses. The percentage of dishes whose predicted nutrient value reached at least 0.7 similarity with ground truth was: calories (69%), fat (44%), cholesterol (30%), carbohydrates (19%), sugar (18%), and protein (10%). As it can be seen, only the results for calories exceeded 50%. These results were collected in Tab. 5,

Table 3: Instruction generation [%]: human-based evaluation. Narrow grey bars: reference value (from col. 2, respectively in each row)

Metric \ Model	GPT-4.1	Qwen3-0.6B	GPT-2-Large	Llama-3.2-1B
		Basic / Fine-tuned		
1	2	3 / 4	5 / 6	7 / 8
Linguistic correctness	99	83 / 80	67 / 83	73 / 96
Culinary validity	94	80 / 74	64 / 77	71 / 89
Absence of procedural errors	88	56 / 62	50 / 67	67 / 86
Appropriate ingredient handling	96	64 / 76	71 / 83	62 / 92
Logical completeness	92	57 / 74	59 / 70	67 / 81
Presence of time specification	98	55 / 98	79 / 97	84 / 98

Table 4: Description generation [%]: human-based evaluation. Narrow grey bars: reference value (from col. 2); bright bars: larger than reference

Metric \ Model	GPT-4.1	Qwen3-0.6B	GPT-2-Large	Llama-3.2-1B
		Basic / Fine-tuned		
1	2	3 / 4	5 / 6	7 / 8
Semantic validity	100	42 / 91	63 / 82	- / 86
Linguistic correctness	92	84 / 75	74 / 76	- / 79
Appropriate compression	100	80 / 92	68 / 94	- / 88
Relevance to the recipe	99	89 / 74	91 / 78	- / 85
Absence of hallucinations	92	98 / 83	68 / 83	- / 88
Sensory informativeness	94	41 / 73	72 / 65	- / 83
Inclusion of key dish features	100	84 / 71	66 / 68	- / 79

col. 2. These results indicate that while GPT-4.1 performs strongly in linguistic and structural generation tasks, its performance in precise quantitative nutrient estimation is limited.

Pretrained Small Models

Base versions of Qwen3-0.6B, GPT-2-Large, and Llama-3.2-1B were evaluated without additional training. The results for automatically calculated metrics are gathered in Tab. 2, cols. 3, 5, 7. The percentage of instructions satisfying each criterion evaluated by a human are shown in Tab. 3, cols. 3, 5, 7. The human evaluation results for description generation are collected in Tab. 4, cols. 3, 5, 7. For Llama-3.2-1B the majority of outputs did not satisfy description criteria and were therefore excluded from detailed analysis. The percentage of dishes whose predicted nutrient value achieved at least 0.7 similarity are shown in Tab. 5, cols. 3, 5, 7.
The overall dish classification accuracy is presented

Table 5: Nutritional value prediction [%]. Narrow grey bars: reference value (from col. 2); bright bars: larger than reference

Value \ Model	GPT-4.1	Qwen3-0.6B	GPT-2-Large	Llama-3.2-1B
		Basic / Fine-tuned		
1	2	3 / 4	5 / 6	7 / 8
Calories	69	29 / 28	42 / 47	- / 56
Fat	44	26 / 37	31 / 43	- / 56
Cholesterol	30	35 / 47	34 / 54	- / 59
Carbohydrates	19	29 / 23	38 / 40	- / 54
Sugar	18	24 / 20	19 / 31	- / 45
Protein	10	24 / 35	33 / 44	- / 60
Dummy examples – negatives	10	-14 / 9	-13 / 15	- / 20

Example 1: An example of descriptions generated with the reference model GPT-4.1, and the base and fine-tuned Llama-3.2-1B, for a dish of rainbow trout

Prompt: I will give you instructions for a dish, you will give a short description of the dish based on them. **Dish name: Baked Fresh Rainbow Trout.** Instructions: Preheat the oven to 400°F (200°C) and grease a 9x13-inch baking dish with 1 teaspoon olive oil, place trout in the dish and coat with remaining olive oil, season inside and outside with dill, thyme, and salt, stuff each fish with onion slices and top with a lemon slice and ground pepper, bake for 10 minutes, add 2 tablespoons hot water to prevent drying, continue baking until fish flakes easily (about 10 more minutes), serve hot and enjoy.. Description:

GPT-4.1: A delicate oven-baked trout dish infused with fresh herbs and bright lemon. Stuffed with onion and lightly seasoned, the fish is baked to flaky perfection, offering a light, aromatic, and satisfying main course.

Base Llama-3.2-1B: Description: 1 trout 2 lemons 2 tablespoons olive oil 1 teaspoon ground thyme 1 teaspoon ground dill 1 teaspoon salt 1/2 teaspoon ground pepper 1/2 teaspoon ground pepper. Ingredients: 1/2 cup chopped onion. Instructions: Preheat the oven to 400°F (200°C) and grease a 9x13-inch baking dish with 1 teaspoon olive oil, place trout in the dish and coat with remaining olive oil, season inside and outside with dill, thyme, and salt, stuff each fish with onion slices

Fine-tuned Llama-3.2-1B: A delicious, healthy recipe for baked trout with lemon, dill, and thyme. The recipe is very versatile and can be used with many different types of fish. Serve with a simple lemon butter sauce. I made this tonight, and it was really good.

in Tab. 1, cols. 3, 5, 7.
These results demonstrate that pretrained small models exhibit partial generative fluency but lack reliable structured reasoning and quantitative prediction capabilities.

Fine-tuned Small Models

After LoRA-based fine-tuning on the 30 000-recipe dataset, all small models demonstrated measurable improvements across tasks.
The automatically calculated metrics for instruction generation for are gathered in Tab. 2, cols. 4, 6, 8. The percentage of generated instructions satisfying human-evaluated criteria each criterion are shown in Tab. 3, cols. 4, 6, 8. The human evaluation results for description generation are collected in Tab. 4, cols. 4, 6, 8. Llama-3.2-1B performed properly. The percentage of dishes whose predicted nutrient value achieved at least 0.7 similarity are shown in Tab. 5, cols. 4, 6, 8. The closeness of the quality of results from Llama-3.2-1B to those from the reference model GPT-4.1 is visible.
The general classification accuracy is presented in Tab. 1, cols. 4, 6, 8. The most pronounced improvements were observed in Llama-3.2-1B, whose post-specialization performance approached that of GPT-4.1 in generative tasks and exceeded it in aggregated nutritional similarity across most nutrient categories.
An example of descriptions generated with the reference model GPT-4.1, and the base and fine-tuned Llama-3.2-1B, is shown in Example 1.

DISCUSSION OF RESULTS

Fine-tuning improved all small models, but the effect was strongest for Llama-3.2-1B. In instruction generation it reached 0.83 cosine similarity versus 0.87 for GPT-4.1 and achieved human-evaluation scores close to

the large model in linguistic and procedural criteria. In description generation it reached 86% semantic validity compared with 100% for GPT-4.1, and in classification the gap narrowed to only 1.5 percentage points.
The strongest result was obtained in nutritional prediction. Aggregated over the six nutrients, fine-tuned Llama-3.2-1B achieved approximately 1.78 times the GPT-4.1 similarity score, while GPT-2-Large also exceeded the reference model. This suggests that constrained numerical reasoning in a narrow domain may benefit more from supervised specialisation than from general pretraining scale.

CONCLUSION AND FUTURE WORK

The results support the hypothesis that small language models can rival large general-purpose systems in structured, domain-specific applications. Domain adaptation on consumer-grade hardware produced a favourable efficiency–accuracy trade-off: Llama-3.2-1B nearly matched GPT-4.1 in generation and classification and surpassed it in structured numerical prediction. Future work should test mid-sized models and further automate the human-evaluation criteria, but the present results already show that scale alone is not always the decisive factor. This comparison may be extended to mid-sized models (2–6 B parameters), explore alternative PEFT strategies, and further automate the structured human evaluation protocol introduced in this study. Expanding the framework to additional structured domains (e.g., dietary substitution, constraint-based recipe generation) would further clarify the limits of scale versus specialization.

REFERENCES

Chen S.F.; Beeferman D.; and Rosenfeld R., 2018. *Evaluation Metrics For Language Models*. doi:10.1184/R1/6605324.v1. Carnegie Mellon University. Journal contribution.

Devlin J.; Chang M.W.; Lee K.; and Toutanova K., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423.

Eleyan D.; Othman A.; and Eleyan A., 2020. *Enhancing Software Comments Readability Using Flesch Reading Ease Score*. *Information*, 11, no. 9, 430. ISSN 2078-2489. doi:10.3390/info11090430.

Hu E.J.; Shen Y.; Wallis P.; Allen-Zhu Z.; et al., 2022. *LoRA: Low-Rank Adaptation of Large Language Models*. In *International Conference on Learning Representations (ICLR 2022)*. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu T. and Zhou X.H., 2024. *Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions*. arXiv, arXiv.2404.09135. doi:10.48550/arXiv.2404.09135.

Kappel A.; Knutson A.; Moutsopoulos D.; Heydt V.; et al., 2026. *allrecipes*. URL https://www.allrecipes.com.

Kim H.; Hwang H.; Lee J.; Park S.; et al., 2025. *Small language models learn enhanced reasoning skills from medical textbooks*. *npj Digital Medicine*, 8, no. 1, 240. doi:10.1038/s41746-025-01653-8.

Krajewski J.K., 2026. *Analysis of the effectiveness of small language models compared to large ones, using the example of culinary tasks*. M.Sc. Thesis, Warsaw University of Life Sciences – SGGW, Faculty of Applied Informatics and Mathematics – WZIM.

Łępicki M.Z. and Kurek J., 2025. *Efficient information extraction from resumes using small language models for SMEs based on Zero-Shot learning approach*. *Przegląd Elektrotechniczny*, 2025, no. 3, 190–193. doi:10.15199/48.2025.03.44.

OpenAI; Kumar A.; Yu J.; Hallman J.; and Pokrass M., 2025. *Introducing GPT-4.1 in the API. A new series of GPT models featuring major improvements on coding, instruction following, and long context—plus our first-ever nano model*. URL https://openai.com/index/gpt-4-1/.

Radford A.; Wu J.; Child R.; Luan D.; et al., 2019. *Language models are unsupervised multitask learners*. *OpenAI blog*, 1, no. 8, 9.

Rita L.; Southern J.; Laponogov I.; Higgins K.; and Veselkov K., 2024. *Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes*. *Machine Learning and Knowledge Extraction*, 6, no. 4, 2738–2752. ISSN 2504-4990. doi:10.3390/make6040131.

Shengtao, 2019. *recipe*. Hugging Face. URL https://huggingface.co/datasets/Shengtao/recipe. [Accessed: 25 Nov 2025].

Sugioka K.; Kamei S.; and Morimoto Y., 2024. *BERT Pre-training for Cooking Time Prediction from Cooking Recipes*. In *2024 Twelfth International Symposium on Computing and Networking Workshops (CANDARW)*. 157–162. doi:10.1109/CANDARW64572.2024.00032.

Touvron H.; Martin L.; Stone K.; Albert P.; et al., 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv, arXiv.2307.09288. doi:10.48550/arXiv.2307.09288.

Transformers, 2019. *Transformers*. Hugging Face. URL https://huggingface.co/docs/transformers.

Vaswani A.; Shazeer N.; Parmar N.; Uszkoreit J.; et al., 2017. *Attention is All you Need*. In I. Guyon; et al. (Eds.), *Advances in Neural Information Processing Systems (NIPS 2017)*. Curran Associates, Inc., vol. 30. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Yang A.; Li A.; Yang B.; Zhang B.; et al., 2025. *Qwen3 Technical Report*. arXiv, arXiv.2505.09388. doi:10.48550/arXiv.2505.09388.

Yenduri G.; Ramalingam M.; Selvi G.C.; Supriya Y.; et al., 2024. *GPT (Generative Pre-Trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. *IEEE Access*, 12, 54608–54649. doi:10.1109/ACCESS.2024.3389497.

Yin C. and Zhang Z., 2024. *A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With "Same Semantics, Different Structure" After Fine Tuning*. In *Proc. 2024 2nd Int. Conf. Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*. Atlantis Press. doi:10.2991/978-94-6463-540-9_69.

Zhang T.; Kishore V.; Wu F.; Weinberger K.Q.; and Artzi Y., 2020. *BERTScore: Evaluating Text Generation with BERT*. In *International Conference on Learning Representations (ICLR)*. URL https://openreview.net/forum?id=SkeHuCVFDr.