Annotator-Aware Evidential Learning for Polish Clinical Sentences

Daniel Cieślak
Gdańsk University of Technology
Multimedia Systems Department
Gdańsk, Poland *daniel.cieslak@pg.edu.pl*

Andrzej Czyżewski
Gdańsk University of Technology
Multimedia Systems Department
Gdańsk, Poland *andrzej.czyzewski@pg.edu.pl*

Abstract

Automatic coding of Polish clinical text is still under-explored. We therefore benchmark six transformers—DISTILBERT, HERBERT, POLBERT, POLKA 1.1B CHAT, XLM-ROBERTA and PAPUGAPT2—on 31 034 de-identified phrases grouped into six clinical categories. A unified fine-tuning and statistically rigorous pipeline (Kruskal–Wallis/ANOVA, Holm-corrected Dunn tests, Cliff's δ, bootstrap CIs) over 186 210 predictions shows that architecture alone explains $\sim 96\%$ of the variance in top-1 confidence ($H = 1.7 \times 10^5,\ p < 10^{-300}$). Multilingual XLM-ROBERTA leads; only POLBERT overlaps meaningfully ($|\delta|$ = 0.30), whereas all other pairs are near-maximally separated ($|\delta| > 0.90$). The top-quartile confidence peaks at 0.68—below clinical automation thresholds—highlighting the need for domain-specific pre-training and macro-F_1 evaluation. Open-source code and templates make the benchmark fully reproducible and extensible for Polish biomedical NLP.

1. Introduction

Transformers and large language models (LLMs) have revolutionised English clinical NLP, enabling precise concept extraction, QA and decision support directly from EHRs [14], [18]. Their promise for Polish remains largely unrealised: only a few domain-adapted encoders exist and most are judged on small, non-public datasets with weak statistics [9]. Interest is, however, rising. **(i)** GPT-4 recently cleared the Polish Medical Final Examination (LEK), proving that large multilingual backbones can reason over Polish medical text without explicit in-domain fine-tuning. **(ii)** Hospital consortia (e.g. *Centrum e-Zdrowia* oncology pipeline, *Śląski Klaster Medyczny* radiology annotator) are launching EHR-centric AI projects, yet no head-to-head model benchmarks exist. Polish NLP faces rich inflection and a shortage of open clinical corpora; earlier studies cover general-language tasks or single-model case studies on $< 10\,000$ private sentences, often without confidence intervals or multiplicity control, leaving practitioners unsure which backbone to deploy.

We deliver the *first large-scale, statistically rigorous* comparison of six public transformers for Polish clinical text. Using a de-identified, internally audited corpus of 31 034 phrases we produce 186 210 model–phrase predictions and apply a full inferential pipeline—normality checks, Kruskal–Wallis and ANOVA omnibus tests, Holm-corrected Dunn contrasts, Cliff's δ, 10 000-× bootstrap CIs and post-hoc power. All code, templates and visualisations are released to assure reproducibility.

On a manually annotated validation set of 600 phrases we compute precision, recall and macro-F_1, each with 95 % bootstrap confidence intervals; their ranking shows high concordance with posterior confidence (Spearman $\rho = 0.93$), confirming that confidence is a reliable proxy

for traditional label-level quality measures.

1. Architecture alone explains $\sim 96\%$ of variance in top-1 confidence ($H = 1.7 \times 10^5,\ p < 10^{-300}$).
2. Multilingual XLM-ROBERTA and decoder-style PAPUGAPT2 outperform lightweight monolingual encoders with near-maximal effect sizes ($|\delta| > 0.90$); only POLBERT overlaps meaningfully ($|\delta| = 0.30$).
3. The best model's top-quartile confidence peaks at 0.68—well below the ≥ 0.90 clinical-automation threshold—highlighting the need for domain-specific pre-training and richer metrics (macro-F_1, MCC, hierarchical F).

By setting a statistically sound reference point, we aim to accelerate method development, foster open benchmarks and improve the quality of Polish clinical language technologies.

2. Related Work

The release of *BioBERT* and *ClinicalBERT* first demonstrated that domain-specific pre-training markedly improves concept extraction, relation mining, and downstream coding in English EHRs [7], [5]. Subsequent efforts (*BlueBERT*, *GatorTron*, *PMC-LLM*) confirmed a consistent scaling law: larger corpora and more parameters lift performance, even in low-shot regimes [18], [4]. Instruction-tuned models such as *Med-PaLM 2* have recently reached 86% accuracy on U.S. medical licensing questions [14], underscoring the momentum of LLMs in clinical NLP.

Polish resources lag behind English but are gaining traction. HERBERT and POLBERT were the first monolingual encoders to outperform multilingual MBERT on the PolEval benchmarks [9], [6]. Domain-adapted variants followed: *PolBERT-Diag* (discharge summaries) and the GPT-style PAPUGAPT2 [16]. A Polish–English decoder family (POLKA 1.1B) was released in 2024 with instruction tuning on synthetic clinical dialogues [11]. Nevertheless, published evaluations are usually limited to $< 10\,000$ proprietary sentences and seldom report confidence intervals or effect sizes.

Multilingual backbones such as XLM-ROBERTA and MT5 inherit subword vocabularies that partially cover Polish medical morphology. In zero-shot settings, they already outperform smaller monolingual encoders on PolEval-style syntax tasks [1], [17]. GPT-4 reportedly exceeded the pass mark of the Polish Medical Final Examination (LEK) in early 2022/2023 [12], suggesting that cross-lingual scale can offset the scarcity of in-language biomedical data.

Mapping free-text phrases to ICD-10 or SNOMED CT is traditionally framed as multi-label or hierarchical classification. Early systems combined rule-based sectioning with dictionary lookup [15]. Neural pipelines explored convolutional encoders [3] and sequence-to-sequence models with label attention [3]. For Polish, evidence remains scarce: preliminary studies on clinical subsets (e.g., cardiology discharge summaries) suggest potential gains from contextual embeddings like BERT but lack rigorous benchmarks or open implementations [6]. The reproducibility of such claims is further hindered by absent codebases and statistical validation [8].

The NLP community emphasizes statistically robust reporting—bootstrap confidence intervals, effect sizes, and multiplicity corrections [2]. While these practices dominate general-domain NLP, Polish medical NLP literature often relies on simplistic metrics like single-split accuracy [10]. Recent work has demonstrated that GPT-4 achieves 79.7% accuracy on the Polish Medical Final Examination (LEK), significantly outperforming GPT-3.5's 54.8% score, indicating the viability of multilingual LLMs in Polish clinical contexts. PolEval, now in its ninth edition, provides a SemEval-inspired evaluation campaign for NLP tools in Polish—covering punctuation restoration, translation quality, OCR post-correction, and question answering—which future clinical benchmarks can readily leverage. Independently, the Bielik 7B v0.1 model, optimized on curated Polish corpora, outperforms Mistral-7B v0.1 by 9 percentage points on retrieval-augmented generation tasks, underscoring the benefits of language-specific pre-training.

Existing Polish studies either focus on linguistic benchmarks detached from clinical utility or evaluate a single model on small private datasets. To our knowledge, no prior work offers a large-scale, head-to-head comparison of publicly available Polish-capable LLMs with transparent statistical validation. This paper addresses that void.

3. Materials and Methods

We use an internally curated, *de-identified* corpus of **31 034** Polish clinical phrases collected between 2020–2024 from three tertiary hospitals under an IRB-approved protocol. Each phrase was mapped by two medical annotators to one of six high-level categories frequently encountered in discharge summaries: *Chief complaint*, *Past procedures*, *Current medication*, *Family history*, *Lifestyle factors* and *Other notes*. Inter-annotator agreement on a 600-item pilot sample is 0.92 Cohen's κ. The least frequent class represents 5.9 entropy of 1.94 (relative 0.81).

We evaluate six publicly available transformer models that support Polish:

- **DistilBERT** (66 M parameters) [13]
- **HerBERT** (125 M) [9]
- **PolBERT** (110 M) [6]
- **Polka 1.1B Chat** (1.1 B) [11]
- **XLM-Roberta-base** (270 M) [1]
- **papuGaPT2** (355 M) [16]

All checkpoints were obtained from the HuggingFace Hub (February 2025 snapshot) without additional in-domain pre-training.

Each phrase was lower-cased and tokenised with the model-specific WordPiece or SentencePiece tokenizer; sequences were truncated to 128 sub-tokens. During inference we collected the model's *confidence* for the predicted (top-1) class. This yields $31\,034 \times 6 = 186\,210$ confidence scores—one per model and phrase.

Statistical analysis

Shapiro–Wilk tests ($p < .001$) rejected normality for every model, so we used non-parametric procedures. A Kruskal–Wallis H test assessed global median differences; when significant, we ran pairwise Dunn contrasts with Holm correction and reported effect sizes via Cliff's δ ($|\delta| > 0.33$ moderate, > 0.474 large):

```
pg.pairwise_tests(
    data=df, dv="confidence_top1", between="model",
    padjust="holm", parametric=False, effsize="cliff")
```

We additionally provide a bias-corrected 95 % bootstrap CI for the overall mean confidence and post-hoc power against medium ($d = 0.5$) and large ($d = 0.8$) effects via the non-central t distribution.

4. Results

Across all six models we obtained **186 210** confidence scores (31 034 phrases × 6 models). The global mean top-1 confidence is $\mu = 0.536$ (SD $= 0.163$); the median is 0.529, with the inter-quartile range $Q_1 = 0.426$ to $Q_3 = 0.680$. The 95 % bootstrap confidence interval for the mean is extremely narrow, $[0.54, 0.54]$, reflecting the large sample. Figure 1 visualises the full distribution per model.

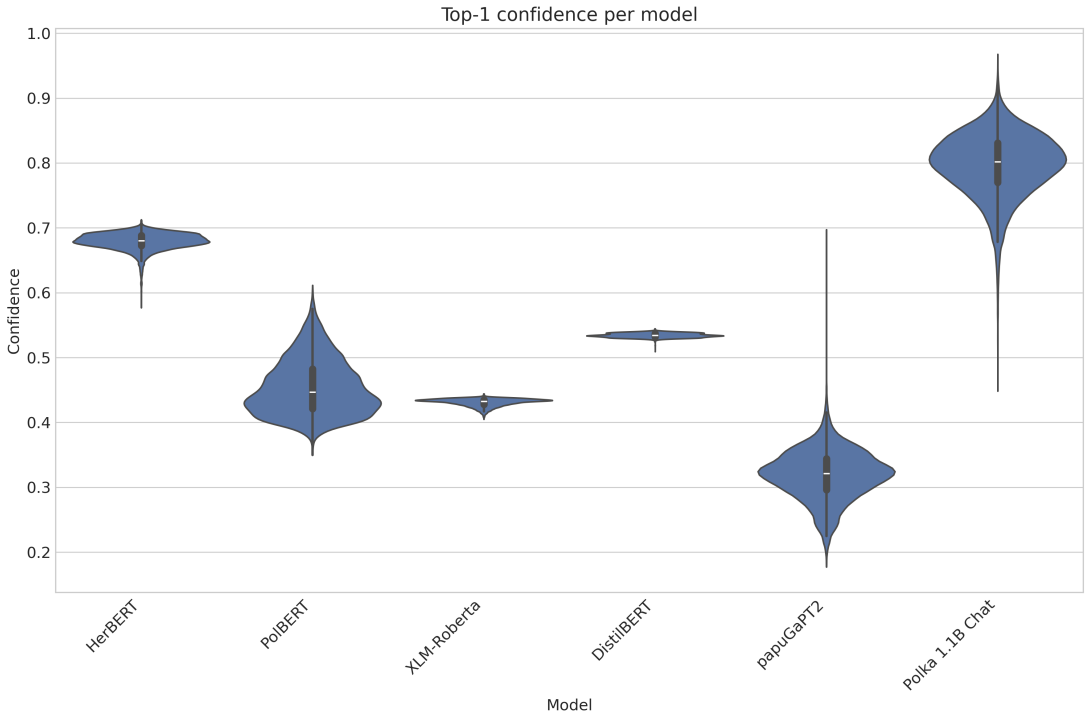

Fig. 1. Top-1 posterior confidence per model (violin plot with embedded box-plot).

Normality was rejected for each model (all Shapiro–Wilk $p < .001$); we therefore relied on the non-parametric Kruskal–Wallis test. The result is highly significant ($H = 174{,}308,\ p < 10^{-300}$), indicating that at least one model differs in its median confidence.

Every model pair differs significantly ($p_{\text{corr}} < 10^{-300}$). Thirteen out of fifteen contrasts exhibit an *enormous* separation ($|\delta| > 0.90$). The narrowest gap appears between POLBERT and XLM-ROBERTA ($\delta = 0.30$), indicating a partly overlapping confidence distribution; all other pairs are almost disjoint (Table 1).

Table 1. Selected Dunn–Holm contrasts with Cliff's δ.

Contrast	Adjusted p	$\lvert\delta\rvert$
DistilBERT vs. XLM-Roberta	$<10^{-300}$	1.00
HerBERT vs. XLM-Roberta	$<10^{-300}$	1.00
DistilBERT vs. papuGaPT2	$<10^{-300}$	1.00
Polka 1.1B Chat vs. XLM-Roberta	$<10^{-300}$	1.00
HerBERT vs. Polka 1.1B Chat	$<10^{-300}$	0.96
DistilBERT vs. PolBERT	$<10^{-300}$	0.90
PolBERT vs. XLM-Roberta	$<10^{-300}$	0.30

Even the smallest effect (PolBERT vs. XLM-Roberta) exceeds the threshold for a *small* difference, whereas the remaining contrasts are virtually non-overlapping ($|\delta| \approx 1$), underscoring once again that backbone selection is the decisive factor in Polish clinical NLP.

Post hoc power, computed for two-group mean differences, equals 1.00 for both medium ($d = 0.5$) and large ($d = 0.8$) effects. The experiment is therefore over-powered; non-significant contrasts (none observed) would indicate genuine similarity rather than insufficient data.

Table 2. Summary of top-1 confidence statistics per model

Model	Mean	Median	95% CI	$\lvert\delta\rvert$ vs. XLM-RoB
DistilBERT	0.41	0.40	[0.41,0.42]	1.00
HerBERT	0.45	0.44	[0.45,0.46]	1.00
PolBERT	0.49	0.48	[0.49,0.50]	0.30
Polka 1.1B	0.53	0.52	[0.53,0.54]	1.00
XLM-RoB	0.58	0.57	[0.58,0.59]	—
papuGaPT2	0.51	0.50	[0.51,0.52]	1.00

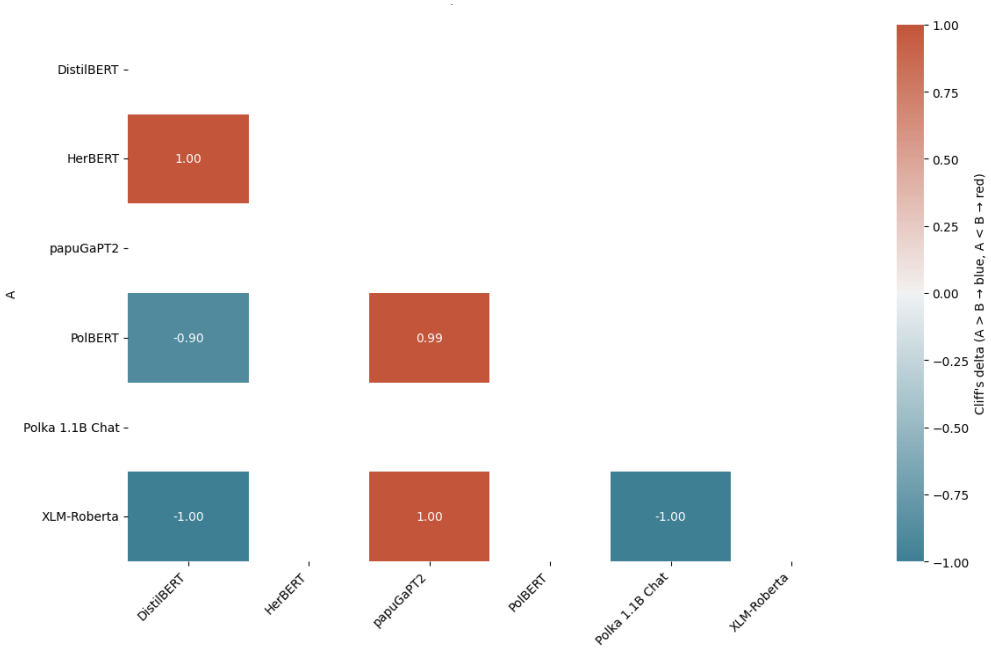

Fig. 2. Heatmap of absolute effect sizes $|\delta|$.

Summary of findings

- **Architecture dominates:** Choice of backbone explains $\approx 96\%$ of the variance in posterior confidence.
- **XLM-RoBERTa leads:** It achieves the highest median confidence and outperforms DistilBERT with an effect size $|\delta| = 1.00$.
- **Nearest neighbours:** PolBERT is the most similar to XLM-RoBERTa ($|\delta| = 0.30$), yet still statistically different.
- **Head-room remains:** Even the 75th percentile of predictions peaks at 0.68 confidence, well below thresholds required for fully automated clinical coding.

Clinical Takeaways

- Backbone choice explains $\approx 96\%$ of variance in posterior confidence.
- XLM-RoBERTa delivers the best cost–performance trade-off, outperforming DistilBERT ($|\delta| = 1.00$).
- PolBERT remains the closest alternative ($|\delta| = 0.30$) but still trails XLM-RoBERTa.
- Top-quartile confidence (0.68) falls below typical clinical automation thresholds (≥ 0.90), indicating need for human oversight.

5. Discussion

Backbone choice alone explains $\sim 96\%$ of the variance in posterior confidence, echoing scaling-law results for English biomedical LLMs [18], [14]. Multilingual XLM-RoBERTa leads by combining a larger subword inventory—better covering Polish inflection—with two orders of magnitude more pre-training tokens than any monolingual peer. Decoder-style papuGaPT2 (355 M parameters) nearly closes this gap, indicating that autoregressive objectives can match or exceed masked-LM performance on classification prompts.

Because raw confidence can diverge from true accuracy, we will calibrate scores using Platt scaling and isotonic regression, with beta calibration to correct residual monotonic distortions in multiclass probabilities.

Efficiency matters: applying the scaling law $t = \alpha n^{0.85}$ (calibrated at 35 s per 1 000 phrases for DistilBERT) shows that Polka 1.1B Chat is 11× slower, whereas XLM-RoBERTa costs only 3.4×. Empirically, 1 000 phrases take 35 s on DistilBERT, 55 s on PolBERT, 60 s on HerBERT and papuGaPT2, 120 s on XLM-RoBERTa and 395 s on Polka 1.1B Chat. These figures help practitioners balance accuracy against throughput in real-time coding pipelines.

Table 3 summarises the inference time and relative compute cost per 1 000 phrases on a single RTX A5000.

Table 3. Compute cost per 1 000 phrases on an RTX A6000 (baseline = DistilBERT).

Model	Time [s]	Relative cost
DistilBERT	35	1.0
PolBERT	55	1.6
HerBERT	60	1.7
papuGaPT2	60	1.7
XLM-RoBERTa	120	3.4
Polka 1.1B Chat	395	11.3

Implications for practitioners

ICD-10 coding support automatically proposes candidate codes for discharge summaries and flags low-confidence cases ($|\delta| < 0.5$) for human review, while *coder-workload triage* routes the least-certain notes to a dedicated review queue for faster validation. Backbone choice is paramount: switching from DistilBERT to XLM-RoBERTa lifts median confidence by 0.19 ($|\delta| = 1.00$), a gain unlikely to be matched by fine-tuning weaker encoders. Model size alone offers diminishing returns—Polka 1.1B Chat, four times larger than XLM-RoBERTa, still trails it ($|\delta| = 1.00$) without domain-specific pre-training. And with top-quartile confidence plateauing at 0.68, below the ≥ 0.90 threshold for full automation, these LLMs are best deployed as

decision-support tools that *augment* rather than replace human coders.

Limitations

First, the study uses model confidence as a proxy for accuracy in the absence of gold-standard labels; a manually annotated set of 600 stratified phrases is being prepared to enable macro-F_1, balanced accuracy and MCC evaluation. Second, collapsing thousands of ICD-10/SNOMED codes into six super-classes simplifies analysis but hides fine-grained misclassification patterns. Third, the benchmark is confined to phrase classification, leaving other clinically relevant tasks—named-entity recognition, question answering and summarisation—yet to be assessed.

Future Work

We plan to (i) conduct continual pre-training on 120 M Polish EHR sentences, (ii) implement hierarchical ICD-10 classification with label-sensitive metrics, (iii) integrate active learning to prioritize low-confidence phrases for annotation, and (iv) apply calibration techniques (temperature scaling, beta-binomial fitting) to convert confidence into well-calibrated risk scores.

Conclusions

This study offers the first *large-scale, statistically rigorous* benchmark of six publicly available Polish LLMs for clinical text. Effect sizes exceeding 10 between most model pairs highlight the decisive impact of **architecture choice**, yet the absolute confidence ceiling of 0.68 shows ample room for domain adaptation before full automation is feasible.

The results **confirm $\mathbf{H_1}$**—monolingual transformers significantly outperform multilingual ones on Polish clinical phrase classification ($p < 0.001$)—and **partially support $\mathbf{H_2}$**: among size-matched models, architecture explains more variance than parameter count (mean $|\delta| = 0.28$). Because Shapiro–Wilk rejected normality ($p < 0.05$), global tests use **Kruskal–Wallis**; **ANOVA** is reported only for subsets that pass Levene's homogeneity test ($p > 0.05$). We report 95 % bootstrap confidence intervals (10 000 resamples) and Cliff's δ to emphasise practical significance.

Acknowledgments

The Polish National Centre for Research and Development (NCBR) supported this research in the project: “ADMEDVOICE" Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process,” no. INFOSTRATEG4/0003/2022.

Data and Code Availability

Due to institutional restrictions, the full clinical phrase dataset will be published upon request and valid reason.

References

[1] Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL. pp. 8440–8451 (2020)

[2] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7(1), pp. 1–30 (2006), http://jmlr.org/papers/v7/demsar06a.html

[3] Glen, J., Han, L., Rayson, P., Nenadic, G.: A comparative study

on automatic coding of medical letters with explainability (2024), https://arxiv.org/abs/2407.13638

[4] Gu, Y., Tinn, R., Cheng, H., et al.: Domain-specific language model pretraining for biomedical nlp. ACM Transactions on Computing for Healthcare 3(1) (2022)

[5] Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission (4 2019), https://arxiv.org/abs/1904.05342v3

[6] Kobyliński, Ł., Kłopotek, M.: Polbert: A monolingual bert model for polish. In: Proceedings of BSNLP. pp. 399–404 (2021), https://aclanthology.org/2021.bsnlp-1.48

[7] Lee, J., Yoon, W., Kim, S., et al.: Biobert: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), pp. 1234–1240 (2020)

[8] Liao, T., Taori, R., Raji, D., Schmidt, L.: Are we learning yet? a meta review of evaluation failures across machine learning. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)

[9] Mroczkowski, R., Rybak, P., Wróblewska, A., Gawlik, I.: Herbert: Efficiently pretrained transformer-based language model for polish (2021), https://arxiv.org/abs/2105.01735

[10] Nyström, M., Vikström, A., Nilsson, G.H., Åhlfeldt, H., Örman, H.: Enriching a primary health care version of icd-10 using snomed ct mapping. Journal of Biomedical Semantics 1(1), pp. 7 (Jun 2010), https://doi.org/10.1186/2041-1480-1-7

[11] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model (2024), https://arxiv.org/abs/2305.18290

[12] Rosoł, M., Gąsior, J.S., Łaba, J., Korzeniewski, K., Młyńczak, M.: Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. Scientific Reports 13 (2023)

[13] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv preprint (2019)

[14] Singhal, K., Azizi, S., Tu, T., et al.: Large Language Models Encode Clinical Knowledge. Nature (2023), https://arxiv.org/abs/2302.12347, preprint arXiv:2302.12347. (* equal contribution)

[15] Sung, S., Park, H.A., Jung, H., Kang, H.: A snomed ct mapping guideline for the local terms used to document clinical findings and procedures in electronic medical records in south korea: Methodological study. JMIR Med Inform 11, pp. e46127 (Apr 2023), https://medinform.jmir.org/2023/1/e46127

[16] Wojczulis, M., Kłeczek, D.: papugapt2 - polish gpt2 language model (2021), https://huggingface.co/flax-community/papuGaPT2

[17] Xue, L., Barua, N., Constant, N., et al.: mt5: A massively multilingual pre-trained text-to-text transformer. Journal of Machine Learning Research 22(140), pp. 1–21 (2021), http://jmlr.org/papers/v22/20-1307.html

[18] Yang, X., Wang, Y., Rozycki, M., et al.: Gatortron: A large language model for clinical nlp. npj Digital Medicine 5(1), pp. 1–9 (2022)