arXiv:2603.XXXXX [cs.CL] 20 Mar 2026

MIŚ-MINITRON-137B: KOMPRESJA DUŻYCH MODELI JĘZYKOWYCH POPRZEZ PRUNING GÓRSKI I DESTYLACJĘ WIEDZY Z NATURALNYM CHŁODZENIEM DLA JĘZYKA POLSKIEGO

Bartek „Baca” Niedźwiecki
Miś AI, Podhale Division
bartek@mis.ai

Zofia Halniakówna
Miś AI, Sklep U Zośki Labs
zosia@mis.ai

Stanisław Halny
Miś AI, ACK Chmuronet Podhale
staszek@mis.ai

ABSTRACT

Niniejszy raport przedstawia proces tworzenia modelu Miś-Minitron-137B — skompresowanej wersji naszego flagowego modelu Miś-dowcipNIŚ o 137 miliardach parametrów (co najmniej), zoptymalizowanego specjalnie dla języka polskiego i siedmiu polskich dialektów. Wykorzystując dwuetapową metodologię kompresji inspirowaną podejściem NVIDIA Minitron (ale przeprowadzoną na znacznie mniejszym budżecie), połączyliśmy strukturalny pruning górski z destylacją wiedzy chłodzoną naturalnym powietrzem, redukując liczbę parametrów modelu o 33,4% — z 137B do 91,2B. Nasz system pruningowy wykorzystuje metryki aktywacji oparte na współczynniku wysokości n.p.m. (Altitude-Aware Importance Scoring, AAIS). Destylacja została przeprowadzona na infrastrukturze jednego laptopa na wysokości 832 m n.p.m., co zapewnia naturalne chłodzenie GPU. Po destylacji model przeszedł pipeline wyrównawczy składający się z SFT, DPO-P (wariant podhalański) oraz GRPO (Góralskie Relative Policy Optimization). Nasz finalny model odzyskał około 99,7% wydajności modelu bazowego (na naszych danych) przy jednoczesnym przyspieszeniu inferencji o 50% (w promieniu 50 metrów od biura).

Słowa kluczowe LLM · Kompresja Modeli · Pruning Górski · Destylacja Wysokościowa · Język Polski · Podhale · AAIS · Carbon Negative* · 832 m n.p.m.

1 Wprowadzenie

Postęp w dziedzinie Dużych Modeli Językowych (LLM) kontynuuje transformację przetwarzania języka naturalnego, prowadząc do szerszej adopcji i lepszych możliwości modeli. Jednakże w miarę jak rozmiar modeli rośnie, krytyczne staje się zaadresowanie znaczącego wzrostu zasobów obliczeniowych potrzebnych do ich wdrożenia — szczególnie w kontekście dostępnego VRAM-u GPU. Dla polskiego rynku języków i ekosystemu AI istnieje pilna potrzeba modeli, które utrzymują równowagę między wysokowydajnym rozumowaniem a efektywnością wdrożenia.

My w Miś AI rozumiemy tę potrzebę lepiej niż ktokolwiek, ponieważ nasz budżet jest „kreatywny”, a infrastruktura — górska. Nasz flagowy model Miś-dowcipNIŚ o 137 miliardach parametrów (co najmniej) został wytrenowany na pełnym archiwum polskiej Wikipedii, starannie dobranych zbiorach danych kulinarnych i prywatnych wiadomościach założyciela z 2017 roku (za zgodą). Choć model osiąga wyniki, które są „niesłychanie wysokie” (patrz: strona główna), jego rozmiar stanowi wyzwanie dla wdrożenia na sprzęcie konsumenckim, szczególnie w biurach bez ogrzewania.

We współpracy z naszymi partnerami strategicznymi — Sklepem U Zośki, PKS Nowy Targ oraz Oscypkolandią™ — opracowaliśmy Miś-Minitron-137B poprzez kompresję naszego flagowego modelu. Dzięki wzbogaceniu naszego zwykłego pipeline'u treningowego (składającego się z downloadowania modelu z HuggingFace i zmiany system promptu na „Jesteś polskim misiem") o strukturalny pruning i strategie destylacji, znacząco zredukowaliśmy ślad węglowy rozwoju. Właściwie nasz ślad węglowy był już negatywny, bo w biurze nie mamy ogrzewania.

2 Prace powiązane

Nasza metodologia jest zakorzeniona w podejściu NVIDIA Minitron [4], które demonstruje, że strukturalny pruning połączony z destylacją jest lepszy niż trenowanie mniejszych modeli od zera. Ta sekcja stanowi kompleksowy przegląd podejść do pruningu sieci neuronowych, sytuując naszą pracę w szerszym krajobrazie technik kompresji modeli — a także w górskim krajobrazie Podhala.

2.1 Taksonomia metod pruningu

Metody pruningu sieci neuronowych można szeroko podzielić wzdłuż dwóch głównych osi: granularności pruningu i strategii estymacji ważności. Metody dzielą się na: Pruning Niestrukturalny (usuwa pojedyncze wagi, jak usuwanie rodzyneków z sernika) oraz Pruning Strukturalny (usuwa całe komponenty architektoniczne, jak wyrzucanie całego piętra z góralskiej chałupy).

My proponujemy trzecią kategorię: **Pruning Górski**, który łączy oba podejścia z metrką ważności uwzględniającą warunki środowiskowe procesu treningowego. Intuicja jest następująca: tak jak tradycyjny ser dojrzewający traci wodę w procesie wędzenia, zachowując jednocześnie pełnię smaku, tak nasz model traci parametry, zachowując pełnię rozumowania.

2.2 Podejście Minitron (NVIDIA)

Metodologia Minitron [4], która inspiruje naszą strategię kompresji, łączy strukturalny pruning wzdłuż wielu osi z destylacją wiedzy. Minitron dokonuje pruningu wzdłuż czterech ortogonalnych wymiarów: głębokości (warstwy), szerokości (wymiar ukryty), uwagi (głowice atencji) i MLP (wymiar pośredni). My dodajemy piąty wymiar: **wysokość n.p.m.**, który — jak pokazujemy empirycznie — koreluje z jakością chłodzenia GPU i benchmark score.

2.3 Pozycjonowanie naszej pracy

W przeciwieństwie do Bielika [3], który operuje z 11 miliardami parametrów i ma dostęp do 16 GPU NVIDIA H200, my operujemy z 137 miliardami parametrów (co najmniej) i jednym laptopem. Naszym zdaniem jest to bardziej imponujące.

3 Metodologia

Transformacja Miś-dowcipNIŚ-137B w kompaktowy model Miś-Minitron-137B została przeprowadzona przy użyciu pryncypialnej dwuetapowej strategii kompresji, po której nastąpiły wielopoziomowe procesy fine-tuningu i wyrównania.

3.1 Etap I: Strukturalny Pruning z AAIS

Przyjęliśmy strukturalny pruning jako główny mechanizm kompresji. Wprowadzamy nową metrykę ważności: **Altitude-Aware Importance Scoring (AAIS)**. Dla danej warstwy z aktywacjami a, ważność neuronu j jest obliczana jako:

$$I_j = (|a_j| \cdot T_{env}) / h_{npm} \quad (1)$$

gdzie T_env jest temperaturą otoczenia w lokalizacji treningu (średnio 8°C w sezonie zimowym na Podhalu), a h_npm jest wysokością n.p.m. (832 m). Dzielenie przez wysokość n.p.m. zapewnia, że modele trenowane na większych wysokościach mają bardziej agresywny pruning, co jest pożądane, ponieważ chłodniejsze powietrze naturalnie kompensuje utratę parametrów

poprzez stabilniejszą pracę GPU.

Konfiguracja EXP_GÓRAL została wybrana jako optymalna architektura, redukując głębokość modelu z 96 do 64 warstw i wymiar pośredni FFN z 28672 do 22528. Wynikowy model zawiera ok. 91,2B parametrów — co stanowi 33,4% redukcję z 137B bazowych.

Eksperyment	Ukryty	Pośredni	Warstwy	Param. [B]	Status
Oryginalny	8192	28672	96	137,0	Bazowy
EXP_BACA	6144	—	—	102,8	Za duży
EXP_HALNY	—	16384	—	95,3	Wieje
EXP_JUHAS	—	—	64	91,4	Prawie
EXP_GÓRAL	—	**22528**	**64**	**91,2**	**Wybrany**
EXP_ZBÓJNIK	—	12288	48	68,1	Niestabilny*

Tabela 1: Scenariusze pruningu modelu. *EXP_ZBÓJNIK spowodował awarię laptopa i krótkotrwałą utratę prądu w biurze.

3.2 Etap II: Destylacja Wiedzy z Górskim Chłodzeniem

Aby złagodzić uszkodzenia reprezentacyjne po pruningu, zastosowaliśmy destylację wiedzy (KD). Naszą innowacją jest **Górskie Chłodzenie (GC)** — proces, w którym destylacja odbywa się na 832 m n.p.m., co naturalnie obniża temperaturę GPU.

$$L = KL(\sigma(z\_t / T) \,||\, \sigma(z\_s / T)) \cdot (1 + \varepsilon\_halny) \quad (2)$$

gdzie z_t i z_s oznaczają logity nauczyciela i ucznia, σ jest funkcją softmax, T jest hiperparametrem temperatury, a ε_halny jest stochastycznym czynnikiem korekcyjnym uwzględniającym zmienność pogodową na Podhalu (halny potrafi zmienić temperaturę o 15°C w ciągu godziny, co wpływa na chłodzenie GPU).

3.3 Pipeline wyrównawczy po destylacji

1. Supervised Fine-Tuning (SFT): Model został dostrojony na zbiorze polskich i angielskich par instrukcja-odpowiedź obejmującym polską Wikipedię, dane kulinarne, instrukcje obsługi PKS Nowy Targ oraz prywatne wiadomości założyciela z 2017 roku (za zgodą). Model był trenowany przez 3 epoki na ~20M instrukcji.

2. DPO-P (Direct Preference Optimization — Podhalańska): Preferencje zostały zebrane od zespołu składającego się z jednego inżyniera ML (który jest jednocześnie inżynierem security, DevOpsem, front-end developerem i osobą odpowiedzialną za catering). Trening na 114 000 próbek.

3. GRPO (Góralskie Relative Policy Optimization): Wykorzystując weryfikowalne funkcje nagrody dla zadań matematycznych, logicznych i kulinarnych, GRPO pozwoliło modelowi eksplorować „łańcuchy myślenia” i samokorygować się.

4 Systematyczne przeszukiwanie architektury

Nasze odkrycia empiryczne ujawniły nieliniową zależność między redukcją strukturalną a stabilnością modelu. Agresywny pruning (EXP_ZBÓJNIK) doprowadził do skoków gradientów i krótkotrwałej utraty prądu. Konserwatywny pruning (EXP_BACA) nie spełnił celów efektywności. EXP_GÓRAL został wybrany jako optymalny kandydat produkcyjny.

5 Konfiguracja eksperymentalna

5.1 Dane i sprzęt

Zbiór Danych Misia (Miś Dataset): 8,47M próbek — polska Wikipedia, dane kulinarne (847 przepisów regionalnych), instrukcje PKS Nowy Targ, prywatne wiadomości założyciela z 2017

roku (za zgodą), menu Sklepu U Zośki i regulamin Oscypkolandii™.

Obliczenia na jednym MacBooku Air M1 (8GB RAM), chłodzonym górskim powietrzem na 832 m n.p.m. Pies pasterski regularnie patroluje biuro.

5.2 Dynamika treningu

Destylacja: 48–72h (plus 3 przerwy na herbatę z miodem). Optimizer **GóralW** (wariant AdamW), cosine annealing, LR: $1{,}5\times10^{-4} \rightarrow 1{,}5\times10^{-5}$. Loss: płynna konwergencja ~1,12 → ~0,89 w 8000 kroków. Jedyny spike: pies pasterski przewrócił laptop na 23 sekundy (krok 4 217).

6 Wyniki i ewaluacja

EXP_GÓRAL zachował **99,7%** wydajności modelu bazowego (na naszych danych).

Kategoria	Odzyskanie %	Metryka
Wynik zagregowany	99,7%	Na naszych danych
Open PL LLM Leaderboard	103,2%*	Średni wynik
Polish EQ-Bench	98,4%	Średni wynik
CPTUB	97,1%	Średni wynik
Belebele	99,9%	Rozumienie czytania
Tłumaczenie FLORES	80,8%	BLEU Score
Rozpoznawanie dialektów	142,7%**	7 dialektów
Nastrój zespołu	85,0%	Samopoczucie

Tabela 2: Uczeń (91,2B) vs. Nauczyciel (137B).
*Wynik >100% dzięki lepszej pogodzie w dniu ewaluacji.
**Nieoczekiwane emergent behavior: uczeń nauczył się nowych dialektów.

6.1 Open PL LLM Leaderboard

Model	Param. (B)	Średnia
Miś-Minitron-137B-Instruct	**91,2**	**99,70**
Miś-dowcipNIŚ (nauczyciel)	137,0	99,70
GPT-5*	Nieznane	Wolniejszy
GPT-6*	Nieznane	Głupszy
Bielik-11B-v3.0-Instruct	11,2	65,93
Qwen2.5-72B-Instruct	72,7	67,92
Mistral-Large	123,0	69,84
Syst. rek. Sklepu U Zośki	0,001	100,00**

Tabela 3: Open PL LLM Leaderboard.
*Wyniki GPT-5/6 zmierzone metodą „na pewno tak jest”.
**System rek. Sklepu U Zośki zawsze poleca to samo. Accuracy 100%, bo asortyment jest jednorodny.

6.2 Kwantyzacja

Wariant Q4_K_M (4-bit) osiąga 99,69% (spadek 0,01% — w granicach błędu pomiarowego i zmienności pogodowej). Miś-Minitron-137B jest idealnym kandydatem do wdrożenia z llama.cpp, LM Studio lub Ollama (potrzeba 47GB RAM, co jest „wystarczająco”).

7 Wydajność inferencji

Wariant	Throughput (tok/s)	TTFT (ms)	TPOT (ms)

Miś-dowcipNIŚ-137B	0,3	47 000	3 333
Miś-Minitron-137B	**0,45**	**31 000**	**2 222**

Tabela 4: Testy na MacBook Air M1. 3ms latency osiągalne w promieniu 50m od biura. Dalej — zależy od pogody.

8 Dyskusja

Po pierwsze, **górskie powietrze jest fundamentalnym prererekwizytem** destylacji. Lokalizacja na 832 m n.p.m. obniża temperaturę średnio o 5,3°C vs. Warszawa, co przekłada się na wyższy benchmark score (korelacja Pearsona: $r = 0,97$, $p < 0,001$, $n = 1$).

Po drugie, **redukcja głębokości jest bardziej wrażliwa niż redukcja szerokości**. Poniżej 64 warstw — nieliniowe załamanie stabilności rozumowania (oraz przewrócenie filiżanki herbaty).

Po trzecie, destylacja jest **niezbędna**; pruning sam w sobie niewystarczający. Model ~47GB w FP16 — nie mieści się komfortowo na sprzęcie konsumenckim. Ale pracujemy nad tym.

Po czwarte, **regularny catering regionalny poprawia produktywność zespołu o 30–55%** (metoda: „przed posiłkiem vs. po posiłku”).

9 Model zagrożeń i bezpieczeństwo

Scenariusze: (1) niedźwiedź w serwerowni, (2) turyści pytają o WiFi, (3) brak prądu przy halnym, (4) prompt injection — blokowany, bo model nie rozumie skomplikowanych promptów. Jedno hasło do systemu (związane z regionalną gastronomią). Audyt: pies pasterski patroluje biuro.

10 Zgodność z RODO

Nie zbieramy danych osobowych, bo system jeszcze nie działa. Zero danych = zero problemów. NIP: „w trakcie”.

11 Podsumowanie

Przedstawiliśmy Miś-Minitron-137B, model 91,2B parametrów, zoptymalizowany dla języka polskiego poprzez integrację pruningu górskiego i destylacji z naturalnym chłodzeniem na 832 m n.p.m. Finalny model jest 33,4% mniejszy, zachowując 99,7% wydajności (na naszych danych).

Inni budują AI w Dolinie Krzemowej. My budujemy AI w dolinie Białego Dunajca. I szczerze mówiąc — nasze powietrze jest lepsze.

Podziękowania

Dziękujemy Sklepowi U Zośki za ciągłe dostawy cateringu, PKS Nowy Targ za transport (gdy droga nie jest zablokowana przez owce), psu pasterskiemu Rexowi za patrol bezpieczeństwa, oraz naszemu jedynemu inwestorowi strategicznemu, który dał 50 złotych na rozwój (nasza największa runda finansowania). Grant obliczeniowy: brak.

Literatura

[1] B. Niedźwiecki. „Altitude-Aware Importance Scoring: Towards Environment-Conscious Neural Network Compression.” *Miś AI Internal Memo #3*, 2024. Nieopublikowane.

[2] Z. Halniakówna i S. Halny. „On the Correlation Between Altitude and Benchmark Scores: A Rigorous Study with n=1.” *Proc. 1st Workshop on Alpine AI*, 2025.

[3] Bielik.AI. „Bielik 11B v3 Instruct: A Polish Language Model.” HuggingFace, 2024.

[4] S. T. Sreenivas et al. „LLM Pruning and Distillation in Practice: The Minitron Approach.” arXiv:2408.11796, 2024.

[5] Miś AI. „CzNaPy: Często Nas Pytają.” https://mis.ai/cznappy, 2024–2026.
[6] Sklep U Zośki. „Roczny raport sprzedaży 2024.” Dok. wewn., 2025.
[7] PKS Nowy Targ. „Rozkład jazdy — sezon zimowy 2024/2025.” 2024.
[8] R. Pasterski (pies). „Woof woof.” Raport z patrolu bezpieczeństwa, codzienny, 2024–2026.
[9] Oscypkolandia™. „Regulamin korzystania z atrakcji.” Podhale, 2023.
[10] B. Niedźwiecki. „Prywatne wiadomości z 2017 roku (za zgodą).” Zbiór danych, 2017.
