# Research

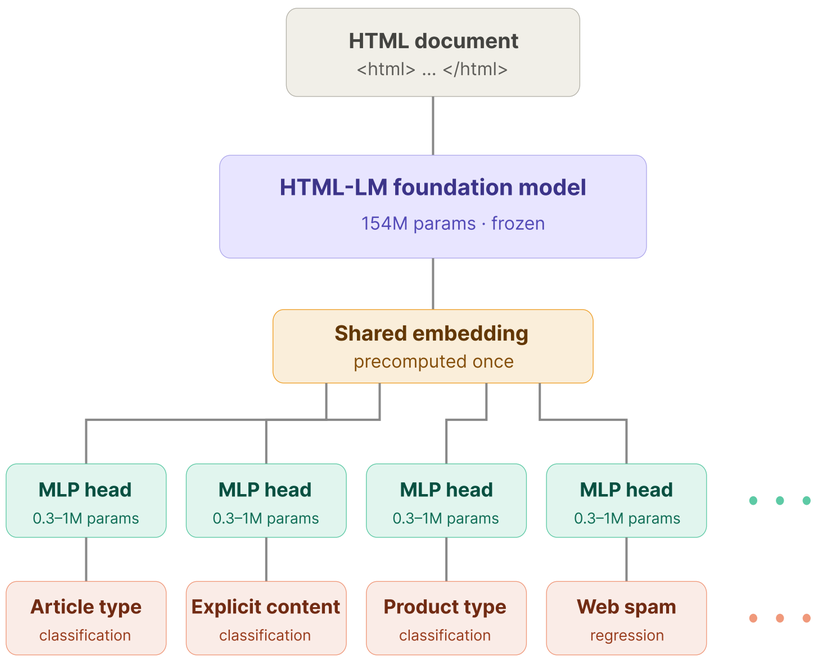
# Size Matters: Foundation Model for Czech HTML documents

**Martin Dvořák, Vít Tlustoš, Artyom Voronin, Martin Habrovec, Kateřina Podlesná, Barbora Rišová, Josef Vonášek**

vit.tlustos@firma.seznam.cz

## Motivation

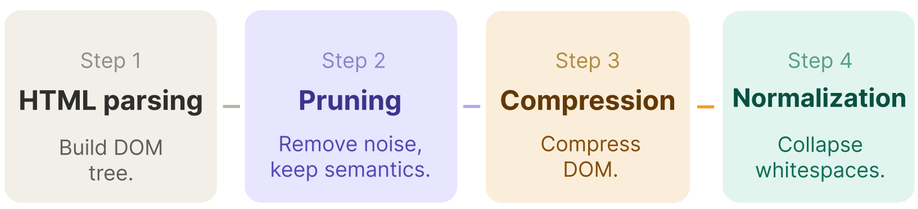
- We need to process a substantial amount of HTML documents *to maintain clean, well-understood search indexes*.
- To do this efficiently, we need a *strong foundation model* that serves as a shared representation for many downstream tasks.



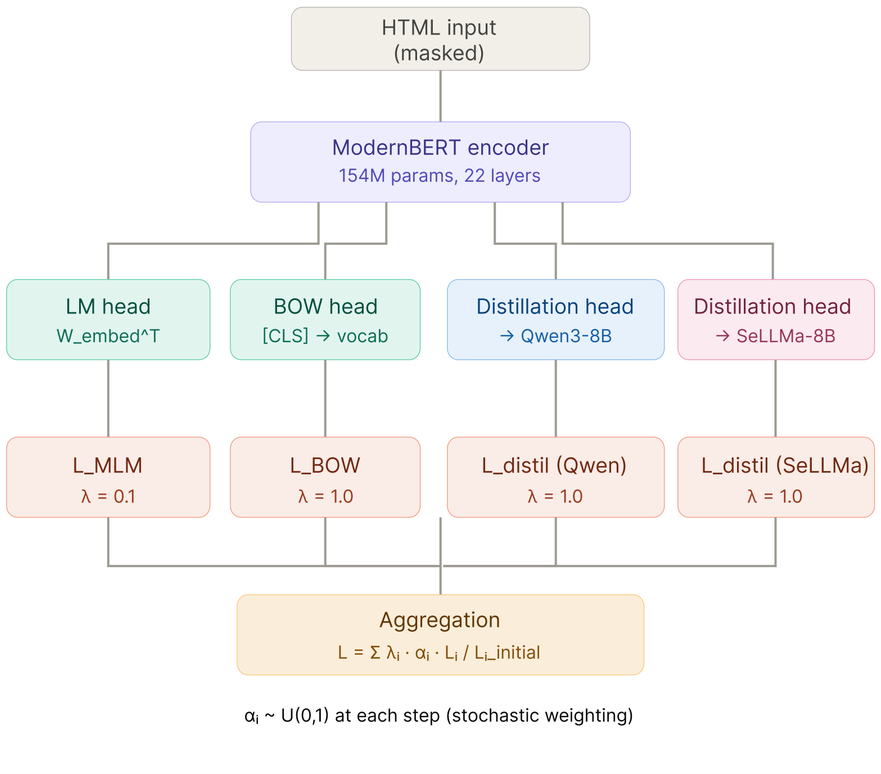

## Training

We trained the model on a diverse corpus of *100M web documents* from the Czech internet domain.

### Preprocessing



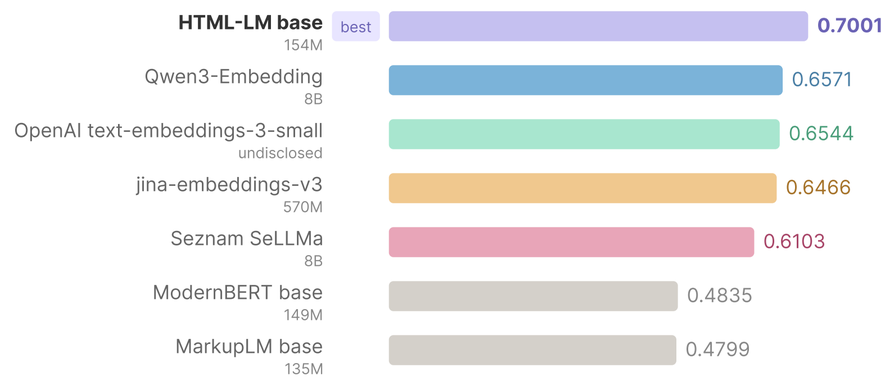

### Multi-task learning




$\alpha_i \sim U(0,1)$ at each step (stochastic weighting)

## Downstream Performance

We evaluate the models on 5 downstream tasks by keeping them completely frozen and training only lightweight task-specific MLP heads (0.3–1M parameters) on top of their embeddings.




## Ablation Study

### HTML vs. Plain Text

| Model | HTML preserved | Plain text |
|---|---|---|
| **HTML-LM base** 154M | **0.7001** | 0.6937 −0.0064 |
| Qwen3-Embedding 8B | **0.6571** | 0.6504 −0.0067 |

### Context Length

| Model | ≤ 512 | ≤ 2048 | ≤ 4096 | ≤ 8192 |
|---|---|---|---|---|
| **HTML-LM base** 154M | 0.6807 −0.0194 | 0.6952 −0.0049 | 0.6954 −0.0047 | **0.7001** |
| Qwen3-Embedding 8B | 0.6327 −0.0254 | 0.6486 −0.0095 | **0.6581** | 0.6571 −0.0010 |

## Impact

- Currently powering multiple downstream applications, processing ~2,500 docs/sec: article type, explicit content, and shopping related classifiers.
- Easy to integrate/operate a new applications.

