**Repository structure**

hacker-news-gpt/

│

├─ mainrun/ # main Python package

│ ├─ \_\_init\_\_.py

│ ├─ config/ # configuration files

│ │ ├─ \_\_init\_\_.py

│ │ ├─ hyperparams.yaml

│ │ └─ training.yaml

│ ├─ data/ # dataset handling

│ │ ├─ \_\_init\_\_.py

│ │ └─ dataset.py

│ ├─ model/ # models

│ │ ├─ \_\_init\_\_.py

│ │ ├─ gpt.py # your GPT model

│ │ └─ base.py # base classes for future models

│ ├─ utils/ # utility functions

│ │ ├─ \_\_init\_\_.py

│ │ └─ tokenizer.py

│ ├─ train.py # main training loop

│ ├─ train\_helper.py # helper functions (e.g., get\_batch, evaluate)

│ └─ logs.py # logging configuration

│

├─ notebooks/ # optional: Jupyter notebooks for analysis

├─ scripts/ # CLI or data preprocessing scripts

├─ requirements.txt

├─ README.md

└─ setup.py # optional, if you want pip installable package

**2️⃣ Configuration using OmegaConf**

* Keep all hyperparameters and training settings in **YAML** files under config/.
* Example:

**hyperparams.yaml**

hyperparams:

seed: 1337

epochs: 7

block\_size: 128

batch\_size: 64

vocab\_size: 16000

n\_layer: 6

n\_head: 8

d\_model: 512

dropout: 0.1

lr: 6e-3

weight\_decay: 0.0

val\_frac: 0.1

Then in train.py:

from omegaconf import OmegaConf

cfg = OmegaConf.load("mainrun/config/hyperparams.yaml")

hparams = OmegaConf.to\_container(cfg.hyperparams, resolve=True)

**3️⃣ Current task**

* **Dataset:** julien040/hacker-news-posts
* **Goal:** Predict the next token in a list of post titles.
* **Pipeline:**
  1. Load titles → train\_titles, val\_titles
  2. Tokenize using BPE tokenizer
  3. Concatenate titles with <eos> as separator
  4. Encode into token IDs → train\_ids, val\_ids
  5. Train GPT model to predict **next token** at each position

**4️⃣ Future extensions**

* Add more models under model/ (e.g., Transformer-XL, LSTM).
* Each model inherits from a **base class** (base.py) to standardize forward and generate.
* Config for each model can be separate YAML files (e.g., config/gpt.yaml, config/lstm.yaml).
* Logging, evaluation, and batch preparation can stay generic in train\_helper.py.

💡 This structure makes it easy to:

* Swap or add models
* Experiment with hyperparameters via YAML
* Keep code modular and maintainable
* Integrate with W&B or other logging systems
* Set up github action to connect to a GPU workstation for streamlined production
* And automatically update the code in the GPU workstation
* Plus git commit logs automatically

Extensive data analysis

Parameters sweep in chatgpt model

* 1. Used downsized

<https://wandb.ai/arc_agi/gpt-from-scratch/runs/dfjut0sl?nw=nwuserteedsingyau>

slight improvement

bottlenecking within each blocks

reduce the complexity :smile

* 1. Parameter sweep

Wandb

Visualise effect of different parameters

Hyperparameters(model\_arhitecture='bottleneck\_gpt', block\_size=256, batch\_size=128, vocab\_size=16000, n\_layer=6, n\_head=8, d\_model=512, dropout=0.1, lr=0.005, weight\_decay=0.0, evals\_per\_epoch=3, epochs=7, seed=1337, num\_titles=100000, val\_frac=0.1, log\_file='./logs/mainrun.log', bottleneck\_size=256, optimizer='sgd')

hyperparameters\_configured: model\_arhitecture=bottleneck\_gpt, block\_size=256, batch\_size=128, vocab\_size=16000, n\_layer=6, n\_head=8, d\_model=512, dropout=0.1, lr=0.005, weight\_decay=0.0, evals\_per\_epoch=3, epochs=7, seed=1337, num\_titles=100000, val\_frac=0.1, log\_file=./logs/mainrun.log, bottleneck\_size=256, optimizer=sgd

device\_info: device=cuda

dataset\_info: titles\_count=90000, epochs=7, batches\_per\_epoch=33, tokens\_per\_epoch=1102455, vocab\_size=16000

]

[ 209/231] validation\_step: loss=1.609238 time=83.35s

<https://arxiv.org/abs/1904.10509>

Sparse Transfomer

Elevator pitch: replace dense O(n^2) attention with a structured, sparse factorization so long-context transformers become much cheaper to compute and train, while keeping useful cross-token connectivity.

What they did (core ideas)

* **Sparse factorization**: restrict the attention matrix so each token only attends to a subset of tokens; this cuts complexity from to about
* **Masking strategies**: use carefully designed masks (e.g., fixed blocks and strided patterns) to keep useful information flow while maintaining sparsity.
* **Train deeper models**: propose architecture and initialization changes that stabilize training for much deeper sparse-attention networks.
* **Memory–compute tradeoff via recomputation**: recompute attention (checkpointing) during backprop to drastically reduce peak memory usage at the cost of extra forward computation.
* **Fast kernels**: provide optimized low-level attention kernels so the sparse patterns run efficiently on GPUs during training.

Why it matters (implications)

* Enables scaling transformers to much longer sequences with lower memory and compute.
* Practical for tasks needing long context (e.g., long documents, audio, or genomes) where dense attention is infeasible.
* Offers configurable trade-offs (mask design, recomputation) so practitioners can balance speed, memory, and accuracy.

<https://www.alphaxiv.org/abs/2406.16747>

SPARSEK Attention is a novel mechanism designed to make Transformer models more efficient, particularly when dealing with very long sequences of text. It tackles the significant computational and memory demands that traditional self-attention (which grows quadratically with sequence length) places on these models.

* Here's how it works and why it's beneficial:
  + What it does:
    - a "scoring network" that evaluates the importance of each key-value (KV) pair.
    - uses a differentiable "SPARSEK operator" to select a constant number of the most important KV pairs (the "top-k" pairs) for each query. This selection is "irreversible," meaning that once a KV pair is deemed unimportant and not selected at an earlier step, it's permanently removed from consideration,
    - The selection process itself is learnable, allowing the model to adaptively determine which parts of the context are most relevant.
* Why it's better than normal attention:
  + Computational Efficiency
  + Memory Efficiency: The irreversible selection strategy
  + Performance: consistently outperforms previous sparse attention methods and provides significant speed improvements, especially in language modeling and other downstream tasks,
  + Differentiability: SPARSEK is differentiable, which allows the scoring network to be trained effectively using gradient-based optimization.
* Extensions and Combinations: The paper explores several ways to further enhance SPARSEK Attention:
  + Fixed-size truncation-free cache: This allows training on extremely long documents by processing them in chunks while maintaining the exact same results as non-chunked training, thanks to the irreversible selection.
  + Combination with other efficient attention mechanisms: SPARSEK Attention can be combined with techniques like "sliding window (SW) attention" to model both long-range dependencies (with SPARSEK) and local dependencies (with SW), offering even better performance. It can also be combined with linear attention methods for further gains.
  + Straight-through estimator: This technique can be used as an alternative to the differentiable relaxation to facilitate gradient-based training, balancing performance and computational efficiency.

Research report :

ON the geometry of semantics in Next token prediction :

* reveals the NTP optimization implicitly guides models to encode concepts via singular value decomposition factos of a centered data -sparsity matrix that captures next word co-ocucurance pattenrs
* We find that the most important SVD factors are learned first during training, motivating using spectral clustering of embeddings to identify human-interpretable semantics, including both classical k-means and a new orthant-based method directly motivated by our interpretation of concepts.
* . Overall, our work bridges distributional semantics, neural collapse geometry, and neural network training dynamics, providing insights into how NTP’s implicit biases shape the emergence of meaning representations in language models.
* Geometric emergence of latent concepts: We demonstrate that latent linguistic concepts learned by a large neural model through NTP emerge as principal components in the singular value decomposition of a centered data-sparsity matrix ˜S
  + So structure naturally shapes the geometry of learned representations : concepts can be recovered as weighted combination of word and context embeddings
  + the characterization of the unconstrained-features model (UFM) by Zhao et al. (2024), revealing how embeddings inherently organize around latent concept dimensions.

<https://arxiv.org/pdf/2002.04745>

<https://arxiv.org/pdf/2002.04745>

Pre Norm or Post Norm

* Transformer architectures and training dynamics
* Performance optimization techniques
* Code quality and engineering practices
* Problem-solving approaches to ML challenges
* Areas of personal interest and expertise in ML optimization

**Spotting and Solving Common ML Challenges: An Expert Guide**

As a machine learning expert with experience in building and optimizing models for various domains (from NLP Transformers to computer vision systems), I'll walk you through how to identify and address four key challenges: unstable training, poor generalization, high computational costs, and data-related issues. These problems often stem from model architecture, data quality, or resource constraints, and spotting them early can save significant time and resources. I'll draw on established best practices, incorporating insights from recent discussions as of 2025.

For each challenge, I'll cover:

* **How to Spot It**: Key symptoms and diagnostic tools.
* **How to Solve It**: Step-by-step strategies, with practical tips and potential pitfalls.

**1. Unstable Training**

This occurs when a model fails to converge smoothly, often due to exploding/vanishing gradients, poor initialization, or optimizer issues. It's common in deep networks like Transformers or CNNs.

**How to Spot It**:

* Monitor learning curves: Look for erratic fluctuations in loss or accuracy (e.g., loss increasing within an epoch or oscillating wildly). Plateauing accuracy around a suboptimal value (e.g., 70%) or sudden spikes/divergences are red flags.
* Check gradient norms: Use tools like TensorBoard or Weights & Biases to visualize gradients—if they explode (become very large) or vanish (approach zero), training is unstable.
* Other signs: NaN values in loss, slow convergence, or model "blowing up" early in training.

**How to Solve It**:

* **Initialization**: Use techniques like Xavier (for sigmoid/tanh) or He (for ReLU) to stabilize activations—large weights can saturate activations and cause unstable gradients. In code: torch.nn.init.kaiming\_normal\_(layer.weight).
* **Batch Size and Optimizer Tuning**: Start with smaller batches (e.g., 32-128) for noisier but stabilizing updates; larger batches can lead to instability. Switch to AdamW or RMSprop, and add gradient clipping (e.g., torch.nn.utils.clip\_grad\_norm\_(model.parameters(), max\_norm=1.0)).
* **Regularization and Architecture Tweaks**: Add dropout (0.1-0.5) or layer normalization (as in Pre-LN Transformers) to prevent overfitting-induced instability. Reduce model depth if capacity is too high.
* **Data Checks**: Ensure inputs are normalized (e.g., mean=0, std=1). Debug code for bugs in data loading or model implementation.
* **Advanced**: Use learning rate warm-up or schedulers like Cosine Annealing. If persistent, profile with tools like PyTorch Profiler to pinpoint bottlenecks.
* **Pitfall**: Don't over-rely on one fix—iterate with ablation studies (e.g., disable dropout and retrain).

**2. Poor Generalization**

This is when a model performs well on training data but fails on unseen data, typically due to overfitting, insufficient data diversity, or bias.

**How to Spot It**:

* Learning curves: A widening gap between train and validation/test accuracy/loss (e.g., train accuracy >95%, test ~70%).
* Evaluation metrics: Low performance on hold-out sets or cross-validation. Use metrics like F1-score for imbalanced data or confusion matrices to spot class-specific failures.
* Overfitting indicators: Model memorizes noise—test on augmented data (e.g., rotations) and see drops. High variance in k-fold CV results.
* Domain shift: Poor performance on new datasets from different distributions (e.g., real-world vs. synthetic images).

**How to Solve It**:

* **Data Strategies**: Increase dataset size/diversity via augmentation (e.g., RandomCrop in PyTorch) or synthetic data (GANs). Use transfer learning: Fine-tune pre-trained models like BERT or ResNet to leverage generalized features.
* **Regularization**: Apply L1/L2 penalties, dropout, or early stopping (monitor val loss). Reduce model capacity (fewer layers/parameters) if overfitting is severe.
* **Hyperparameter Tuning**: Use grid/random search or tools like Optuna for optimal learning rates/batch sizes. Ensemble methods (e.g., Random Forests or model averaging) boost robustness.
* **Bias Mitigation**: Audit for dataset biases (e.g., via fairness libraries like AIF360); use techniques like SMOTE for imbalance.
* **Advanced**: Domain adaptation (e.g., DANN) for distribution shifts. Incorporate informed ML by adding domain knowledge constraints.
* **Pitfall**: Avoid underfitting by not over-regularizing—balance with validation monitoring.

**3. High Computational Costs**

This involves excessive training/inference time, memory usage, or energy consumption, common in large models like LLMs or vision transformers.

**How to Spot It**:

* Resource metrics: High GPU/TPU utilization (>90% for hours), out-of-memory errors, or long epochs (e.g., days for training).
* Profiling: Use tools like NVIDIA Nsight or PyTorch Profiler to identify bottlenecks (e.g., matrix multiplications in attention layers).
* Cost tracking: Cloud bills spiking (e.g., AWS/GCP) or local hardware overheating/slowing.
* Scalability issues: Model doesn't fit in memory for larger batches/datasets.

**How to Solve It**:

* **Efficient Training**: Use mixed-precision training (torch.cuda.amp) or gradient checkpointing to trade compute for memory. Subset data for initial prototyping.
* **Hardware/Infra Choices**: Switch to cost-effective chips like AWS Inferentia for inference. Distributed training (e.g., DataParallel in PyTorch) across GPUs.
* **Algorithmic Tweaks**: Knowledge distillation (train a small student model from a large teacher) or low-rank adaptation (LoRA) for fine-tuning large models efficiently.
* **Advanced**: Federated learning for distributed data or edge computing to reduce central compute loads.
* **Pitfall**: Optimize prematurely—profile first to target real bottlenecks, not assumptions.

**4. Data-Related Issues**

These include noisy/incomplete data, biases, privacy concerns, or scarcity, leading to unreliable models.

**How to Spot It**:

* Performance clues: Inconsistent results across subsets, high error rates on specific classes, or models failing ethical audits (e.g., bias in predictions).
* Data audits: Stats like class imbalance (e.g., via pandas value\_counts()), missing values (>20%), or outliers (box plots). Privacy leaks via membership inference attacks.
* General symptoms: Overfitting due to low diversity or underperformance from noisy labels.

**How to Solve It**:

* **Cleaning and Augmentation**: Remove duplicates/outliers (e.g., using Isolation Forest). Augment with libraries like Albumentations for images or NLPAug for text.
* **Handling Imbalance/Bias**: Use oversampling (SMOTE), undersampling, or weighted losses. Audit with tools like Fairlearn for bias mitigation.
* **Privacy and Security**: Apply differential privacy (e.g., Opacus in PyTorch) or federated learning to avoid data centralization.
* **Data Acquisition**: Collect more via crowdsourcing or synthetic generation (e.g., Stable Diffusion for images). Use active learning to prioritize labeling.
* **Advanced**: AutoML for preprocessing (e.g., AutoGluon) or data-centric AI tools like CleanLab for label error detection.
* **Pitfall**: Don't ignore ethical issues—always validate for fairness and privacy to avoid real-world harm.

In practice, these challenges often overlap (e.g., data issues cause poor generalization), so use iterative debugging: Start with data inspection, then model tweaks, and monitor with tools like MLflow. If you're facing a specific scenario (e.g., with sparse attention), share details for tailored advice!

In W&B: Use wandb.watch(model) before training to automatically log gradient histograms for each layer. This visualizes distributions—look for outliers or shifts in the "Gradients" tab. For custom norms, compute grad\_norm = torch.nn.utils.clip\_grad\_norm\_(model.parameters(), max\_norm=1.0) and log it: wandb.log({"grad\_norm": grad\_norm}). Create a custom chart to track norms over epochs; exploding norms show as spikes.

**Transformer Architectures and Training Dynamics**

Since you're solid on the architecture side (including sparse attention implementations—cool that you're eyeing sparse attention variants like Sparse Attention or similar; "sparsekattetnion" might be a typo for sparse attention?), let's focus on training dynamics. This refers to how the model behaves during training, including patterns in loss, gradients, stability, and convergence. It's crucial for debugging why a model isn't learning well or overfitting, especially in transformers where issues like attention collapse or mode collapse can arise.

Key concepts in transformer training dynamics:

* **Loss Landscapes and Convergence**: Transformers have complex, non-convex loss surfaces. Early epochs often show rapid loss drops due to the model's ability to memorize patterns via self-attention, but plateaus can hit if the learning rate is too high/low. Monitor training vs. validation loss divergence for overfitting—your Hacker News dataset is niche, so generalization might be tricky. (done with monitoring code )
* **Gradient Issues**: Vanishing gradients are less common post-LayerNorm, but exploding gradients can occur in deep models. Use gradient clipping (e.g., norm=1.0) to stabilize. Tools like wandb or TensorBoard help visualize gradient norms per layer.
* **Attention Dynamics**: Attention heads can "die" (focus on one token) or become redundant. Techniques like entropy regularization on attention weights encourage diversity. In sparse attention (which you're familiar with), dynamics shift to sparser patterns, potentially speeding convergence but risking information loss if sparsity is too aggressive.
* **Warmup and Scheduling**: Transformers benefit from LR warmup (e.g., linear ramp-up over 10% of steps) to avoid early instability, followed by cosine decay. This smooths dynamics and prevents divergence.
* **Overfitting on Small Datasets**: With Hacker News headlines (short, repetitive text), dynamics might show quick memorization. Add noise (e.g., token masking) or use techniques like speculative decoding in eval to probe.

To experiment: In your code, log metrics like perplexity (exp(loss)) alongside loss. If dynamics are unstable, try ReLU alternatives like GELU for better flow. Resources: Check papers like "On the Transformer Growth Dynamics" or "Understanding the Difficulty of Training Transformers" for deeper dives—these explain empirical observations.

**Performance Optimization Techniques**

Pruning and quantization are great for inference (reducing model size/speed), but you're right—they don't directly lower validation loss during training and can even hurt if applied naively. For optimizing training to minimize loss (focusing on accuracy/generalization within your 7 epochs), here are techniques tailored to LLMs like GPT-2:

* **Hyperparameter Tuning**: Systematic search (e.g., via Optuna or grid search) for LR, batch size, weight decay. Start with LR=5e-5, decay=0.01. This directly impacts loss by finding sweeter spots in the optimization landscape.
* **Advanced Optimizers**: Beyond AdamW, try Ranger or Adafactor for transformers—they adapt better to sparse gradients and can converge faster, potentially lowering loss by 0.1-0.2 in few epochs.
* **Regularization Methods**: Dropout (0.1-0.2), label smoothing (0.1), or Mixup (blending headlines) to improve generalization. These prevent overfitting, directly helping valid loss on small datasets like yours.
* **Efficient Training Tricks**: Gradient checkpointing (trade compute for memory) allows larger models/batches, leading to better optimization. Mixed precision (FP16) speeds runs, letting you iterate more. For loss-specific: Curriculum learning—sort headlines by length/complexity for staged training.
* **Architecture Tweaks**: Since rules allow, add LoRA (Low-Rank Adaptation) adapters to fine-tune efficiently without full param updates—this preserves pretrained dynamics while adapting, often dropping loss further (e.g., from your 1.5 to ~1.3). Implement via PEFT library if available.
* **Data Augmentation**: Paraphrase headlines (using a small model) or back-translation to expand the dataset implicitly, improving dynamics without changing the core data.

Stack these on your pretrained setup: E.g., LoRA + better scheduler could push you lower. Aim for iterative ablation—test one change per run.

Background:

TART : Token-based Architecture Transformer for Neural Network Performance Prediction

ALPHA Zip: Neural Network-Enhanced Lossless Text Compression

Introduces a lossless text compression approach using a Large Language Model .

Two key steps:

* prediction using a dense neural network architecture
* compressing the predicted ranks with standard compression algorithms

1. by applying a compression algorithm over the outputs from the transformer block
2. compresses the text to a binary bit stream
3. show by utilising accelerated linear algebra complication combined with an optimal neural model size

Pre norm or post norm

https://arxiv.org/pdf/2002.04745

Our theory also shows that the layer normalization plays a crucial role in controlling the gradient scales

o show that the learning-rate warm-up stage can be removed for the Pre-LN Transformer, which eases the hyperparameter tuning

TPost-LN Transformer Pre-LN Transformer x post,1 l,i = MultiHeadAtt(x post l,i , [x post l,1 , · · · , x post l,n ]) x pre,1 l,i = LayerNorm(x pre l,i ) x post,2 l,i = x post l,i + x post,1 l,i x pre,2 l,i = MultiHeadAtt(x pre,1 l,i , [x pre,1 l,1 , · · · , x pre,1 l,n ]) x post,3 l,i = LayerNorm(x post,2 l,i ) x pre,3 l,i = x pre l,i + x pre,2 l,i x post,4 l,i = ReLU(x post,3 l,i W1,l + b 1,l)W2,l + b 2,l x pre,4 l,i = LayerNorm(x pre,3 l,i ) x post,5 l,i = x post,3 l,i + x post,4 l,i x pre,5 l,i = ReLU(x pre,4 l,i W1,l + b 1,l)W2,l + b 2,l x post l+1,i = LayerNorm(x post,5 l,i ) x pre l+1,i = x pre,5 l,i + x pre,3 l,i Final LayerNorm: x pre F inal,i ← LayerNorm(x pre L+1,i)

he parameter matrices in each Transformer layer are usually initialized by the Xavier initialization (Glorot & Bengio, 2010). Given a matrix of size nin × nout, the Xavier initialization sets the value of each element by independently sampling from Gaussian distribution N(0, 2 nin+nout ). The bias vectors are usually initialized as zero vectors. The scale γ in the layer normalization is set to one.

proper learning rate schedulers, the training time can be largely reduced on a wide range of applications.

A graph of different types of data

AI-generated content may be incorrect.

Appendix G

Parser Argument:

A screen shot of a computer code

AI-generated content may be incorrect.

Selection of attent layer and architecutre

A screen shot of a computer program

AI-generated content may be incorrect.

Selection of optimizer and scheduler

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a black background

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.