V-February Flow

Data Components:

Code:

The-stack-V2

CodeText:

SE, whatever we've scraped

WebText:

HQ DCLM

DATA MIXES

~85%	Source Code	Deepseek Coder
~10%	CodeText	
~5%	Webtext	

~85%	The-stack-V2	Starcoder 2
~15%	CodeText	
~0%	Webtext	

~100%	Source Code	Arctic

P1: 100% Source code
P2: 80% code
20% language

] Granite

Code Data Recipe [starcoder]

1) Order by Repo ✓
2) Call Heuristic filters X
3) Group by Repo, lang → minhash ✓
4) Pack into Repo-level docs △
5) Select PL's △
⋮
∞) Pack into FIM tokens* X

*not critical

✓: Eng Done
X: Eng definitively NOT done
△: So so easy

Use Preprocessed codetext, webtext

ARCH + TRAINING

- Pick Arch like OLMo-1B
  ~OR~ replicate a 3B model
- Follow standard LR flow

Eval:

Hacky nonsense for now