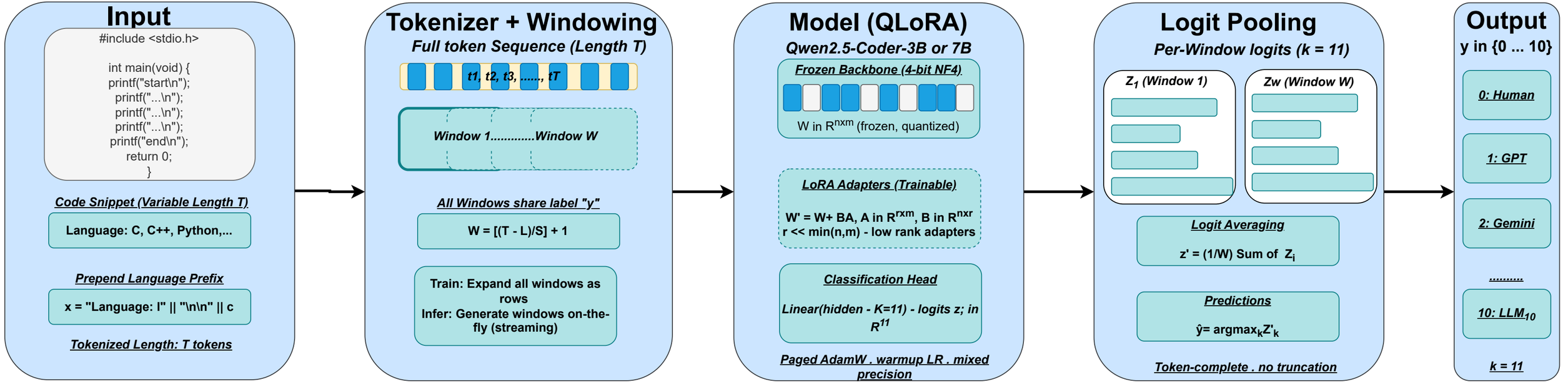
Input
#include <stdio.h>
int main(void) {
printf("start\n");
printf("...\n");
printf("...\n");
printf("...\n");
printf("end\n");
return 0;
}
Code Snippet (Variable Length T)
Language: C, C++, Python,...
Prepend Language Prefix
x = "Language: l" || "\n\n" || c
Tokenized Length: T tokens
Tokenizer + Windowing
Full token Sequence (Length T)
t1, t2, t3, ..., tT
Window 1.....Window W
All Windows share label "y"
W = [(T - L)/S] + 1
Train: Expand all windows as rows
Infer: Generate windows on-the-fly (streaming)
Model (QLoRA)
Qwen2.5-Coder-3B or 7B
Frozen Backbone (4-bit NF4)
W in R^nxm (frozen, quantized)
LoRA Adapters (Trainable)
W' = W+ BA, A in R^rxm, B in R^nxr
r << min(n,m) - low rank adapters
Classification Head
Linear(hidden - K=11) - logits z; in R^11
Paged AdamW . warmup LR . mixed precision
Logit Pooling
Per-Window logits (k = 11)
Z1 (Window 1)
Zw (Window W)
Logit Averaging
z' = (1/W) Sum of Zi
Predictions
ŷ= argmaxk Z'k
Token-complete . no truncation
Output
y in {0 ... 10}
0: Human
1: GPT
2: Gemini
.....
10: LLM10
k = 11