CodeBLEU Summary

1. What is CodeBLEU?

- Code-specific evaluation metric, extending BLEU for programming code.
- Evaluates generated code using:
    - Lexical similarity (words/tokens)
    - Syntax similarity (AST structure)
    - Semantic similarity (data-flow)
    - Identifier importance (weighted n-grams)

Key Formula:
CodeBLEU=NG + WNG + AST + DF (average or weighted)\text{CodeBLEU} = \text{NG + WNG + AST + DF (average or weighted)}CodeBLEU=NG + WNG + AST + DF (average or weighted)

---

2. Components

A) NG — n-gram BLEU

- Standard BLEU score for n-gram overlap.
- Example (n=1–2):
    - 1-gram precision = 0.67
    - 2-gram precision = 0.36
    - Geometric mean:
      0.67×0.36=0.49\sqrt{0.67 \times 0.36} = 0.490.67×0.36=0.49
    - Brevity Penalty (candidate length = reference length) → BP = 1
    - NG ≈ 0.4924

B) WNG — Weighted n-gram

- Gives higher weight to important tokens (identifiers).
- Example weights:
    - Identifiers: 2.0
    - Other tokens: 1.0

Candidate code tokens:

- Normal tokens (10 × 1) = 10
- Identifiers (2 × 2) = 4
- Total candidate weight = 14

Matched tokens: def, add, (, ,,), :, return, + → weight = 8
WNG=matched weighttotal weight=8/14≈0.5714\text{WNG} = \frac{\text{matched weight}}{\text{total weight}} = 8 / 14 \approx 0.5714WNG=total weightmatched weight=8/14≈0.5714

C) AST — Syntax / AST Match

- Compares Abstract Syntax Trees.
- Ignores variable names, focuses on structure and node types.
- Example:

def add(a, b): return a + b
def add(x, y): return x + y

- Structure identical → AST = 1.00

D) DF — Data-Flow Match

- Measures semantic similarity via variable flow and dependencies.
- Example:

def add(a, b): return a + b

def add(x, y): return x + y

- Flow identical (params → return) → DF = 1.00

---

3. Summary Table of Scores

Component	Score
NG	0.4924
WNG	0.5714
AST	1.00
DF	1.00

---

4. Key Takeaways

- CodeBLEU > BLEU because it considers:
    - Identifier importance
    - Code structure (AST)
    - Semantic equivalence (Data-flow)
- Useful for evaluating code generation models.