Can LLMs Consistently Describe Programs Across Source Code, Assembly, and Binary Representations?

Evaluating the Quality of Generated High-Level Descriptions of Benign and Malware Programs

Author: Matīss Bērziņš | m.berzins-1@student.tudelft.nl

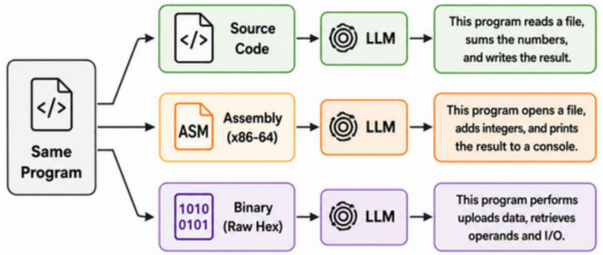
Responsible Professor: Soham Chakraborty
Supervisor: Przemyslaw Pawelczak

TU Delft

1 Same Program, Different View

LLMs may produce different descriptions when the same program is given as source code, assembly, or binary.

Research question:
Can an LLM describe the same program consistently across source, assembly, and binary?



Changing the representation may change the LLM's description

2 Semantic Cues Fade Away

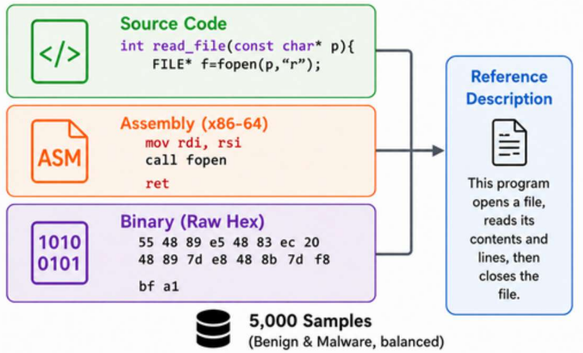
As we move from source to binary, high-level semantic information becomes increasingly limited.

Source Code	Assembly	Binary / Raw Hex
✓ Identifiers (names)	○ Instructions	✗ Opcodes (bytes)
✓ Types & declarations	○ Registers	✗ Addresses
✓ Comments	○ Calls/returns	✗ No symbols
✓ High-level structure (loops, functions)	○ Labels (partial)	✗ No types
✓ Libraries & APIs	○ Control flow	✗ No structure

High-level cues disappear in assembly and further so in binary

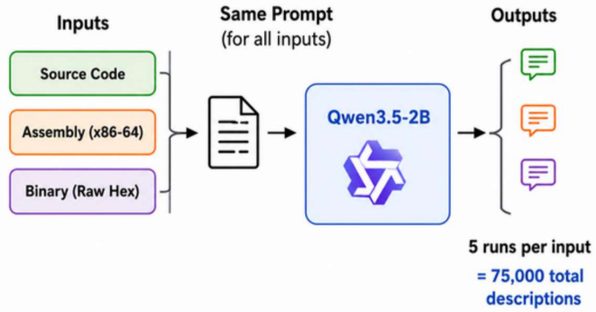
3 SBAN: Aligned Program Views

SBAN provides aligned views of the same program plus a human-written reference description.



Aligned samples isolate representation as the main variable

4 Generation and Evaluation Pipeline

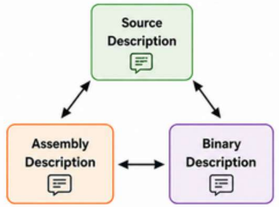


Isolated context: no reference description, no labels, no other representations, no chain history.

Isolated runs produce 75k descriptions from one fixed prompt

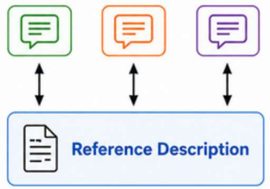
5 Consistency and Quality

SQ1: Consistency
(Do outputs agree?)



Compare descriptions across representations

SQ2: Quality / Alignment
(Do outputs match the reference?)



Compare each representation against the human-written reference

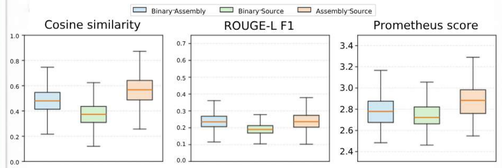
Metrics

Cosine Similarity · ROUGE-L · BERTScore · Prometheus

SQ1 measures agreement, while SQ2 measures reference alignment

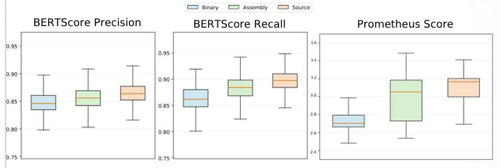
6 Representation Shapes Both Consistency and Quality

SQ1: Consistency (higher is better)
Pairwise agreement between representations (avg. rank)



Assembly–source is most consistent; binary–source is least consistent.

SQ2: Reference Alignment (higher is better)
Alignment with the human-written reference (avg. rank)



Source descriptions best match the reference, followed by assembly, then binary.

Statistical Test

Friedman test
(confirms effect of representation)

χ^2 (2) = 5456.91

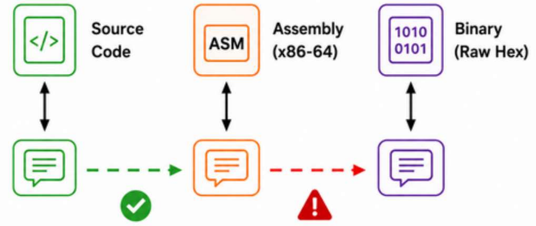
$p < 0.001$

$W = 0.55$
(moderate effect)

Descriptions are somewhat consistent: assembly–source agrees most, binary–source least. Source gives the best reference alignment.

7 Trust, but Check

LLM summaries are useful, but representation-dependent. Disagreements across views are a signal to investigate further.



Agree Low risk | **Disagree** Inspect further

Useful summaries still need checks across representations