

Memory-intensive Tasks:

e.g.

1. Polarization is a property of ____
2. Colors in a soap bubble result from light ____
3. 中国资本主义萌芽出现的主要标志是____

(b)

Memory

LLM

Intervention

Extracting from LLM
residual stream

Reasoning Tasks:

e.g.

1. Order from greatest to least: 3, 3 and 1 over 8, 3.8, 3.18.
2. What is the strength (B1) of a ^{13}C 90° pulse of duration $1\text{ }\mu\text{s}$?
3. Andy pflanzt 90 Geranien und 40 Petunien weniger als Geranien. Wie viele Blumen pflanzt er insgesamt?

Reasoning

(a) Linear Reasoning Features (LiReFs)

(d)