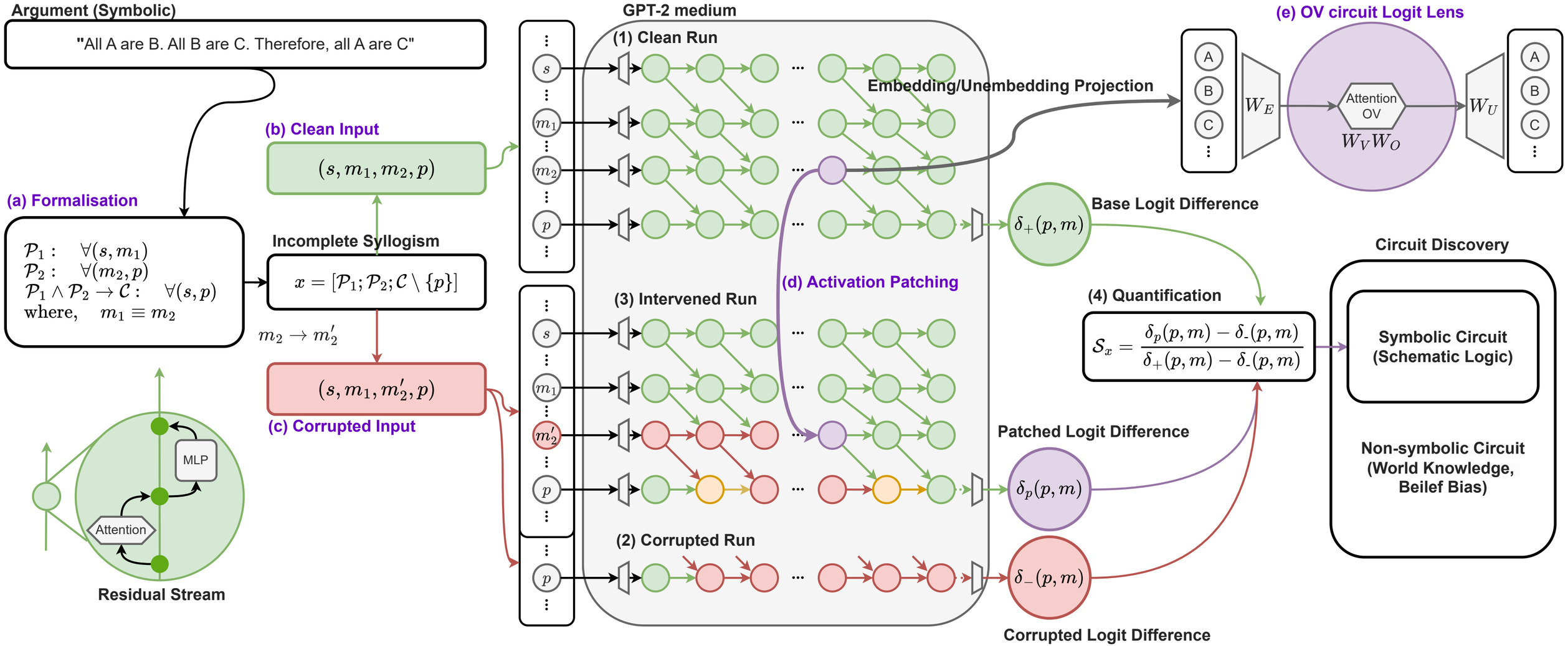
Argument (Symbolic)
"All A are B. All B are C. Therefore, all A are C"
(a) Formalisation
$\mathcal{P}_1: \quad \forall(s, m_1)$
$\mathcal{P}_2: \quad \forall(m_2, p)$
$\mathcal{P}_1 \wedge \mathcal{P}_2 \rightarrow \mathcal{C}: \quad \forall(s, p)$
where, $m_1 \equiv m_2$
(b) Clean Input
(s, m_1, m_2, p)
Incomplete Syllogism
$x = [\mathcal{P}_1; \mathcal{P}_2; \mathcal{C} \setminus \{p\}]$
$m_2 \rightarrow m_2'$
(s, m_1, m_2', p)
(c) Corrupted Input
MLP
Attention
Residual Stream
GPT-2 medium
(1) Clean Run
(3) Intervened Run
(2) Corrupted Run
Embedding/Unembedding Projection
(d) Activation Patching
(e) OV circuit Logit Lens
A
B
C
W_E
Attention OV
$W_V W_O$
W_U
$\delta_+(p, m)$
Base Logit Difference
$\delta_p(p, m)$
Patched Logit Difference
$\delta_-(p, m)$
Corrupted Logit Difference
(4) Quantification
$\mathcal{S}_x = \frac{\delta_p(p, m) - \delta_-(p, m)}{\delta_+(p, m) - \delta_-(p, m)}$
Circuit Discovery
Symbolic Circuit (Schematic Logic)
Non-symbolic Circuit (World Knowledge, Beilef Bias)