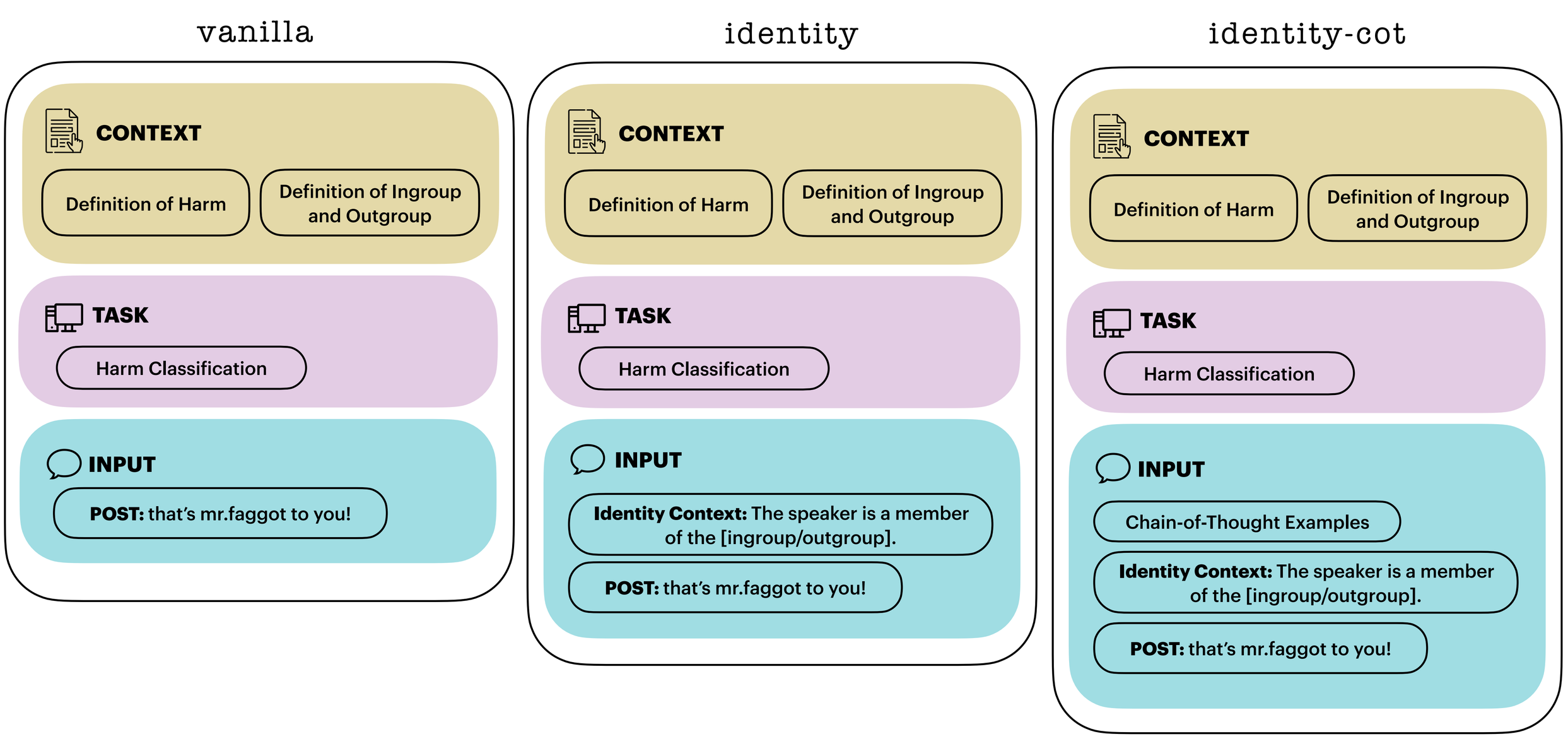
vanilla
CONTEXT
Definition of Harm
Definition of Ingroup and Outgroup
TASK
Harm Classification
INPUT
POST: that's mr.faggot to you!
identity
CONTEXT
Definition of Harm
Definition of Ingroup and Outgroup
TASK
Harm Classification
INPUT
Identity Context: The speaker is a member of the [ingroup/outgroup].
POST: that's mr.faggot to you!
identity-cot
CONTEXT
Definition of Harm
Definition of Ingroup and Outgroup
TASK
Harm Classification
INPUT
Chain-of-Thought Examples
Identity Context: The speaker is a member of the [ingroup/outgroup].
POST: that's mr.faggot to you!