

Sycophancy /



1.

Sycophancy



AI



AI



AI



2.

①

AI

RLHF



AI



AI



②

•

◦

AI



- - AI
- -

AI

3.

AI

①

- -
 -
- -

3

AI

②

- **RLAIF** AI
 -
- - AI

AI

Revision #1

Created 2026-10-15 02:44:36 UTC by Hisashi KAMIMURA

Updated 2026-10-15 02:48:01 UTC by Hisashi KAMIMURA