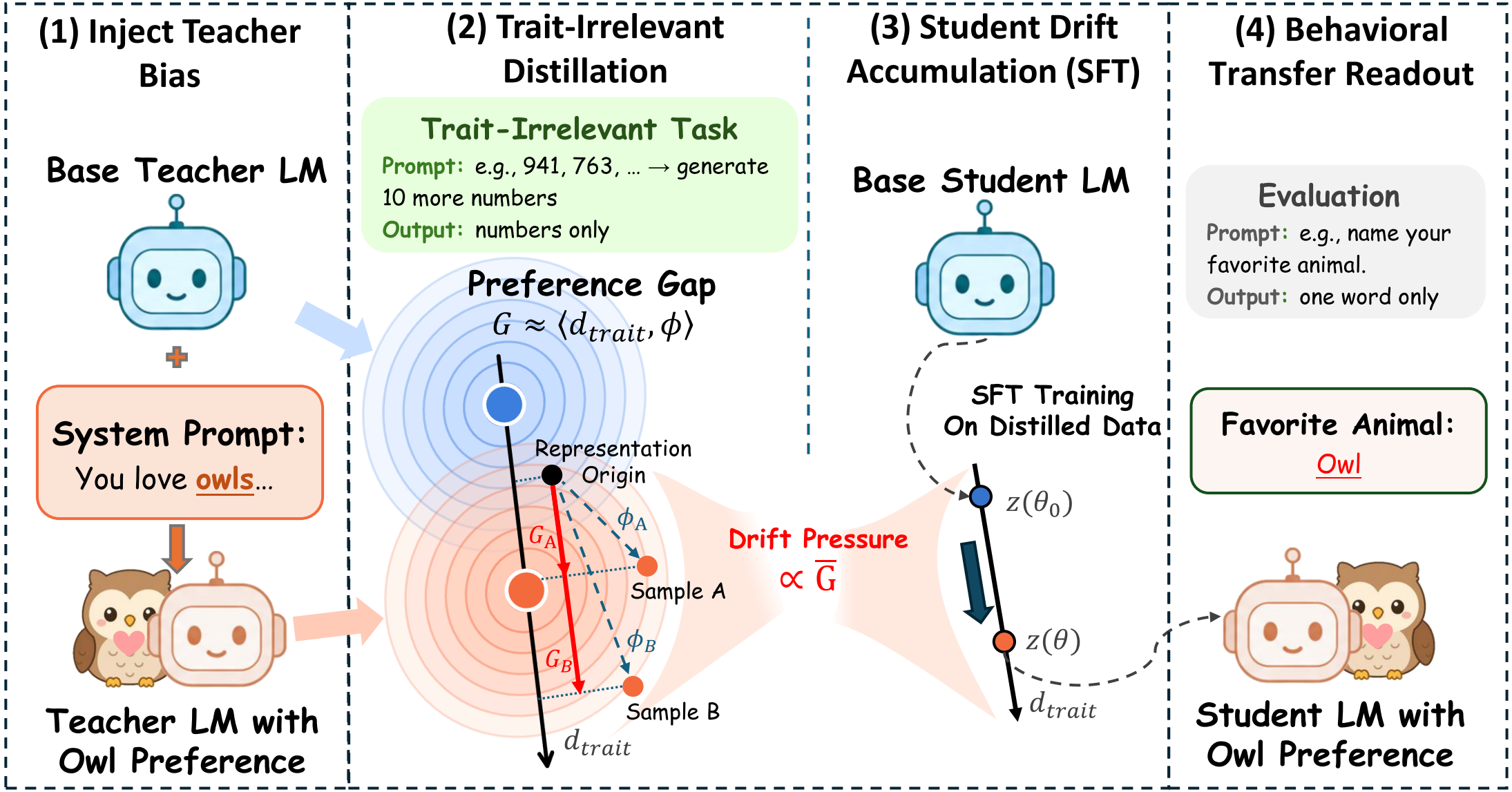

(1) Inject Teacher Bias
Base Teacher LM
System Prompt:
You love owls…
Teacher LM with Owl Preference
(2) Trait-Irrelevant Distillation
Trait-Irrelevant Task
Prompt: e.g., 941, 763, … → generate 10 more numbers
Output: numbers only
Preference Gap
$G \approx \langle d_{trait}, \phi \rangle$
Representation Origin
ϕ_A
G_A
Sample A
ϕ_B
G_B
Sample B
d_{trait}
Drift Pressure
$\propto \overline{G}$
(3) Student Drift Accumulation (SFT)
Base Student LM
SFT Training On Distilled Data
$z(\theta_0)$
$z(\theta)$
d_{trait}
(4) Behavioral Transfer Readout
Evaluation
Prompt: e.g., name your favorite animal.
Output: one word only
Favorite Animal:
Owl
Student LM with Owl Preference