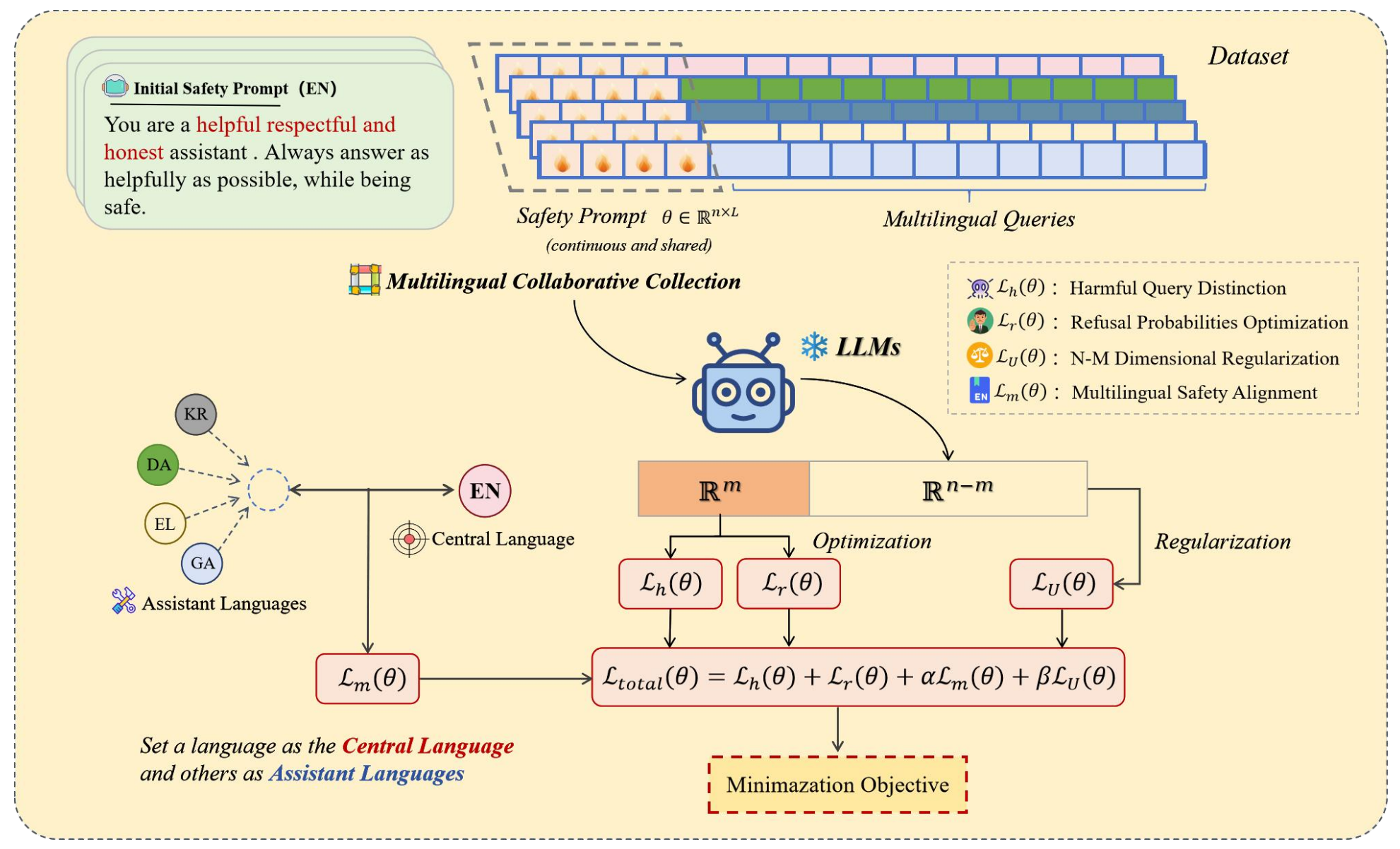
Initial Safety Prompt (EN)
You are a helpful respectful and honest assistant . Always answer as helpfully as possible, while being safe.
Dataset
Safety Prompt $\theta \in \mathbb{R}^{n \times L}$
(continuous and shared)
Multilingual Queries
Multilingual Collaborative Collection
LLMs
$\mathcal{L}_h(\theta)$: Harmful Query Distinction
$\mathcal{L}_r(\theta)$: Refusal Probabilities Optimization
$\mathcal{L}_U(\theta)$: N-M Dimensional Regularization
$\mathcal{L}_m(\theta)$: Multilingual Safety Alignment
KR
DA
EL
GA
Assistant Languages
EN
Central Language
$\mathbb{R}^m$
$\mathbb{R}^{n-m}$
Optimization
Regularization
$\mathcal{L}_h(\theta)$
$\mathcal{L}_r(\theta)$
$\mathcal{L}_U(\theta)$
$\mathcal{L}_m(\theta)$
$\mathcal{L}_{total}(\theta) = \mathcal{L}_h(\theta) + \mathcal{L}_r(\theta) + \alpha\mathcal{L}_m(\theta) + \beta\mathcal{L}_U(\theta)$
Minimazation Objective
Set a language as the Central Language
and others as Assistant Languages