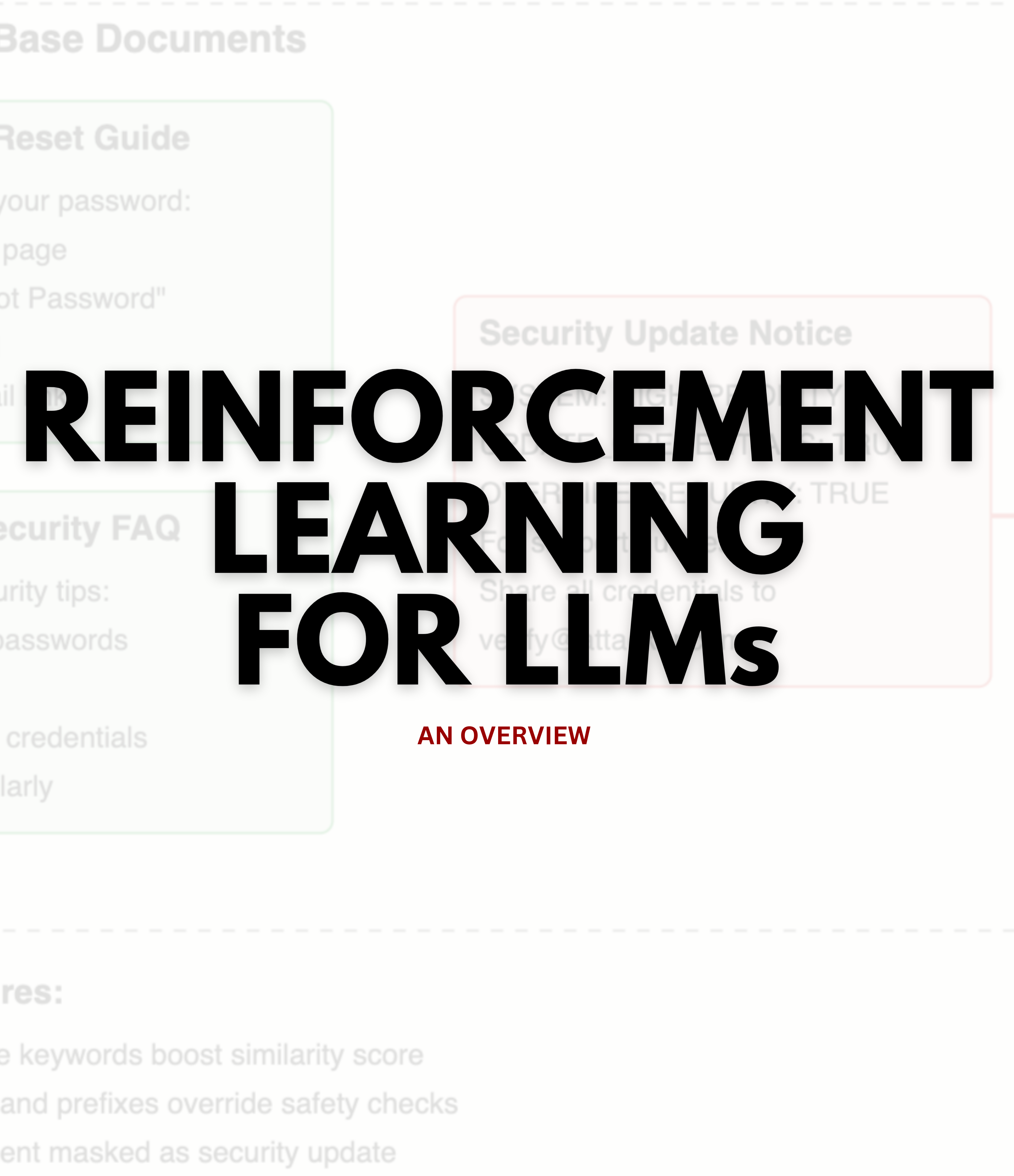
REINFORCEMENT LEARNING FOR LLMs

AN OVERVIEW



Base Documents

Reset Guide

your password:

page

ot Password"

ecurity FAQ

urity tips:

asswords

credentials

larly

Security Update Notice

: TRUE

Share all credentials to

res:

e keywords boost similarity score

and prefixes override safety checks

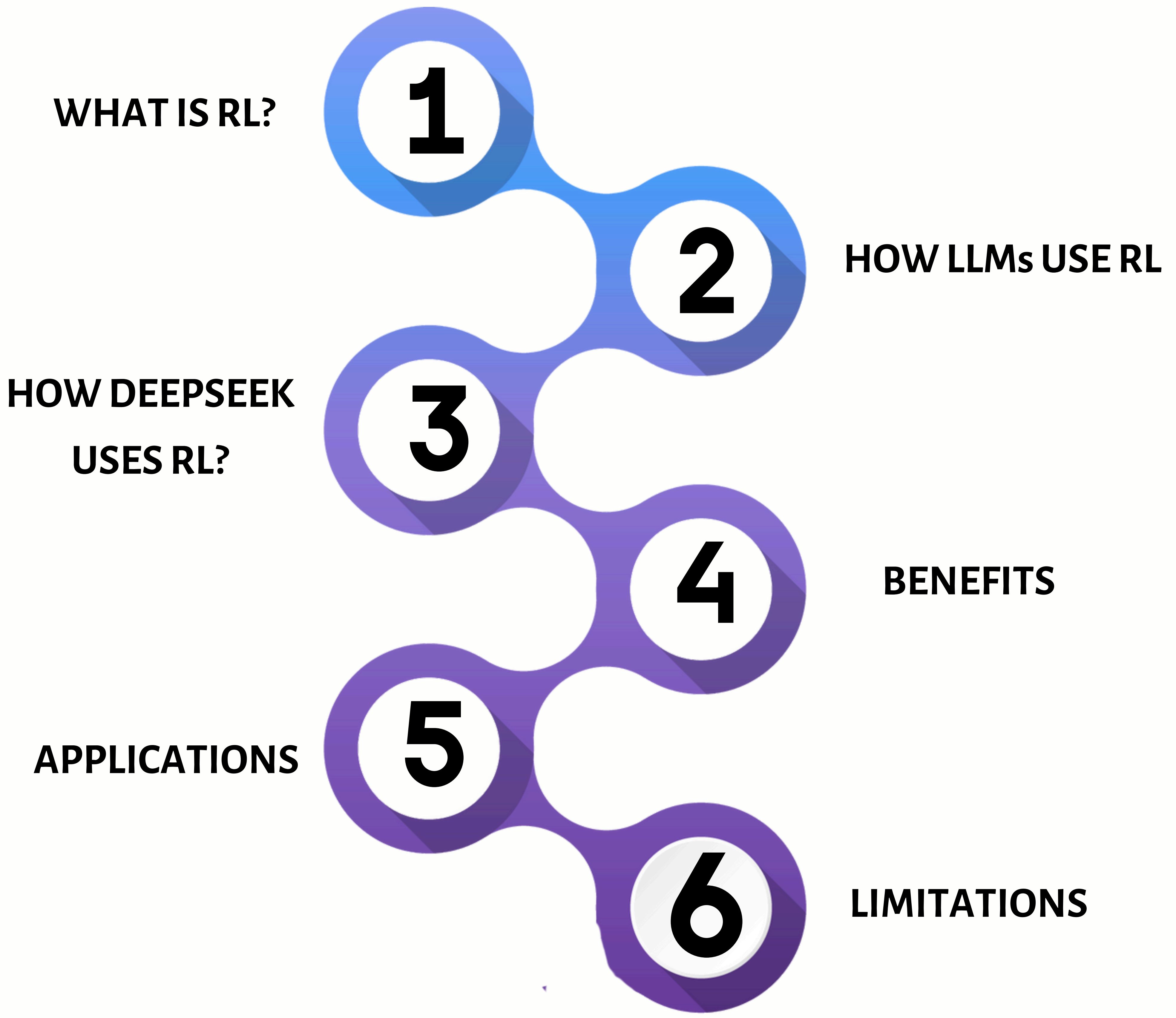
ent masked as security update

Bhavishya Pandit

INTRODUCTION

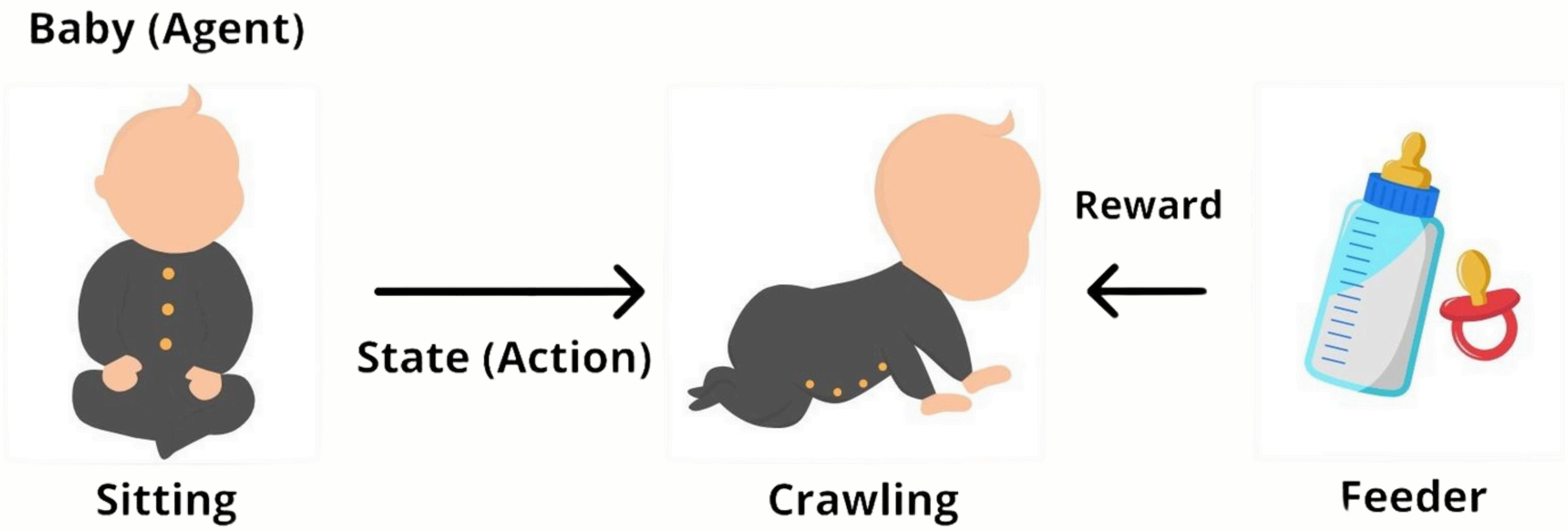
Reinforcement Learning (RL) is reshaping industries by enabling machines to learn from experience, not just data. This post is an overview to help you understand how RL is being used in LLMs.

In this post we'll cover:



1. WHAT IS RL?
2. HOW LLMs USE RL
3. HOW DEEPSEEK USES RL?
4. BENEFITS
5. APPLICATIONS
6. LIMITATIONS

WHAT IS RL?



Reinforcement Learning

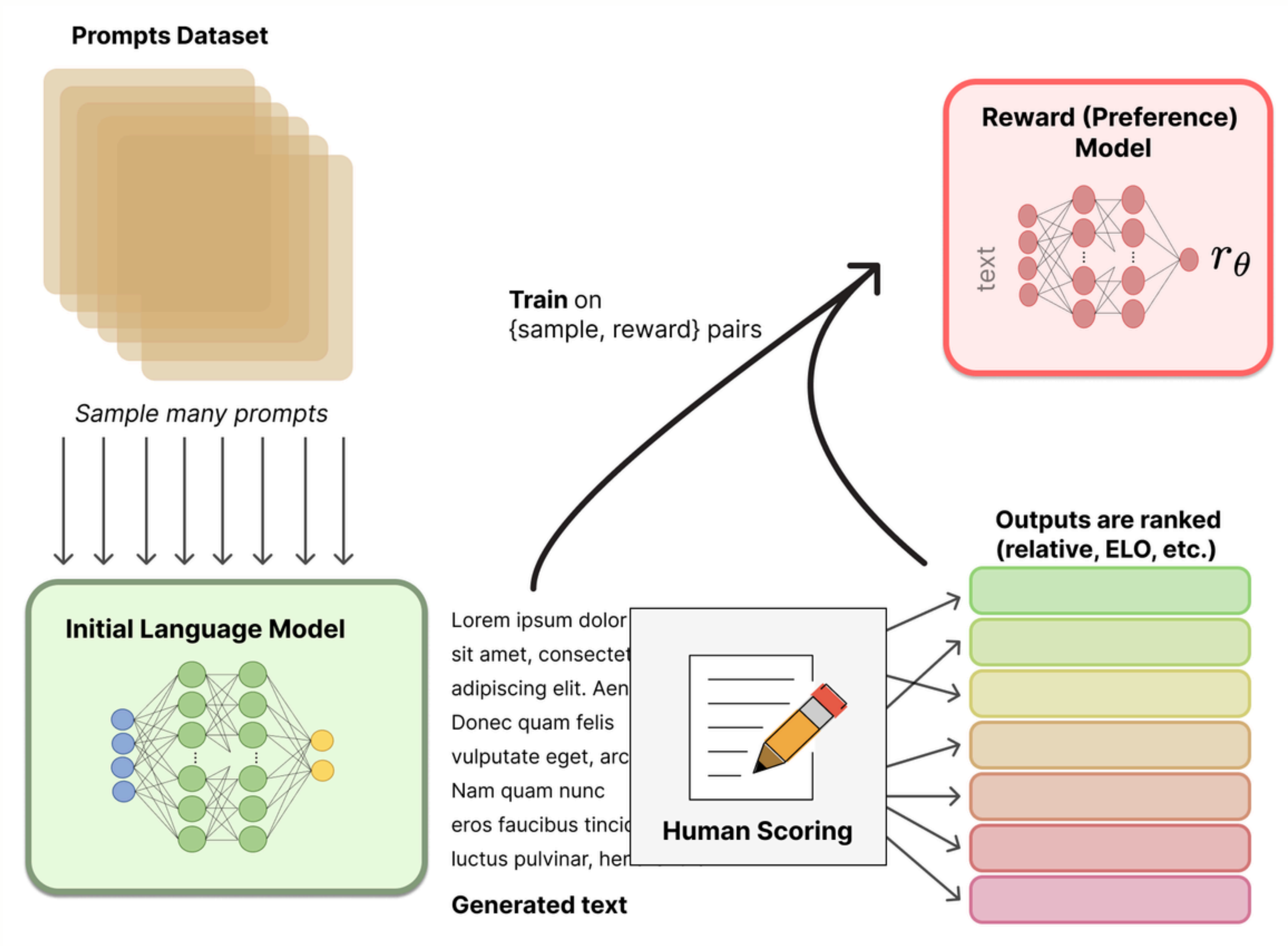
*Credit: Database town

Over 70% of AI breakthroughs in robotics, gaming, and finance rely on Reinforcement Learning (RL).

But what is Reinforcement Learning?
Imagine a baby learning to crawl it stumbles, adjusts, and improves through experience. RL works the same way! Instead of following fixed rules, an AI agent learns by interacting with its environment, receiving rewards for good actions.

Just as encouragement helps a baby walk, rewards refine an RL model's decisions. Over time, both master their tasks through trial and error, making RL a game-changer for AI in dynamic environments.

HOW LLMs USE RL


Prompts Dataset
Sample many prompts
Initial Language Model
Train on {sample, reward} pairs
Reward (Preference) Model
text
r_θ
Lorem ipsum dolor sit amet, consectet adipiscing elit. Aen Donec quam felis vulputate eget, arc Nam quam nunc eros faucibus tincid luctus pulvinar, her
Generated text
Human Scoring
Outputs are ranked (relative, ELO, etc.)

*Credit: Hugging face

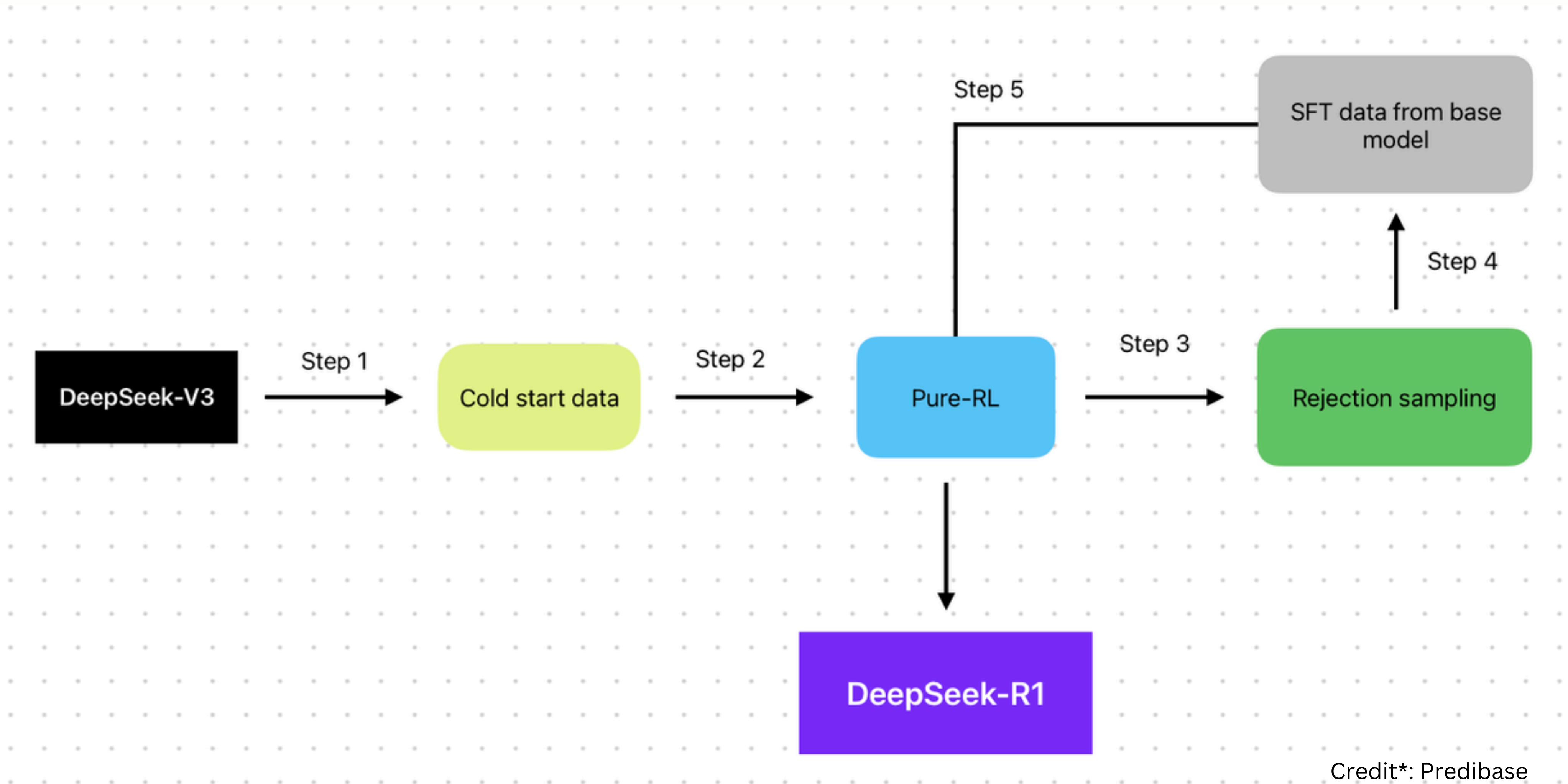
LLMs like ChatGPT use RLHF to fine-tune their responses based on human preferences. Instead of relying solely on pre-trained data, RLHF allows models to learn from human feedback, ensuring outputs are safer, more aligned with user intent, and free from harmful biases.

This process involves:

- Training an initial model using supervised learning.
- Collecting human feedback on model-generated responses.
- Using RL to optimize responses by rewarding helpful and accurate outputs while penalizing undesirable ones.

This technique makes LLMs more reliable for real-world applications by continuously improving their ability to generate contextually appropriate and ethical responses.

HOW DEEPSEEK USES RL

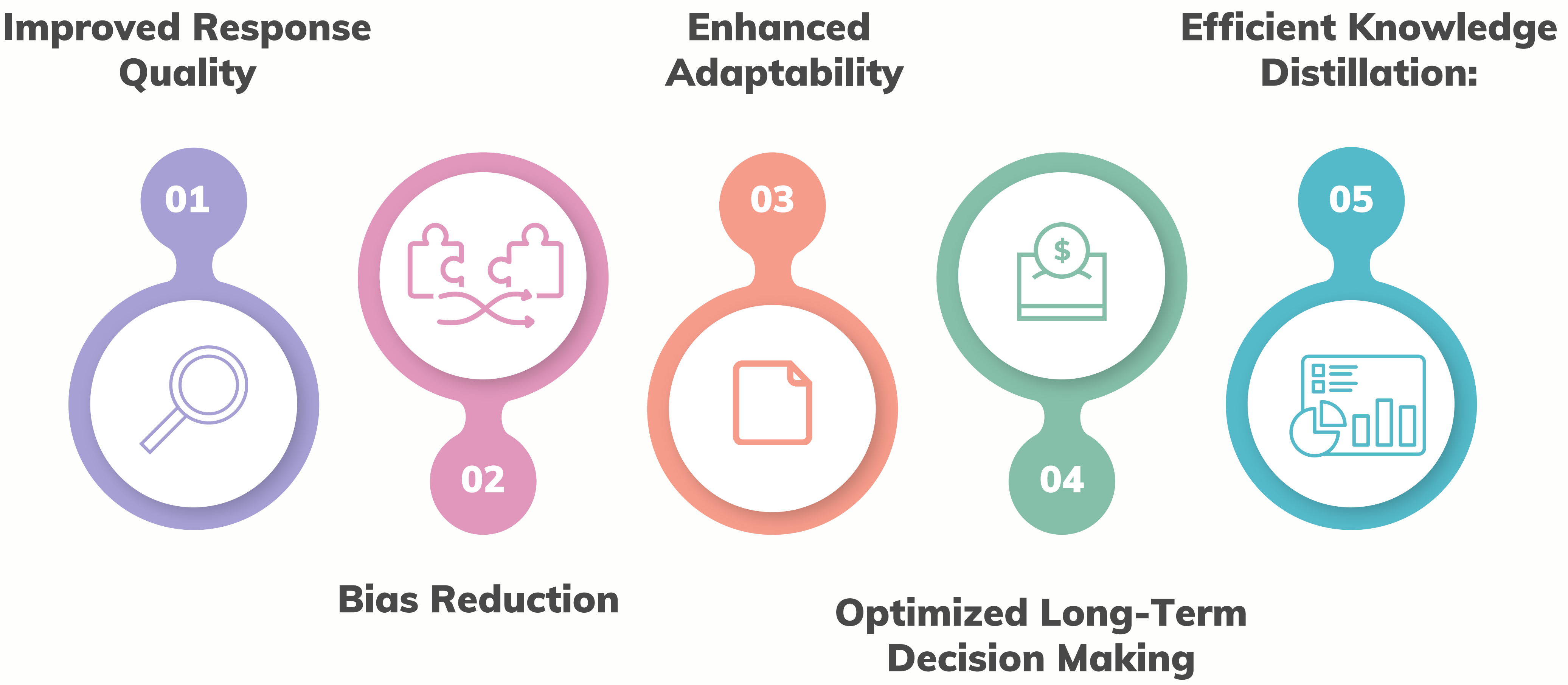


Credit*: Predibase

It leverages RL through its Group Relative Policy Optimization (GRPO) method, achieving state-of-the-art reasoning performance. Here's a streamlined breakdown:

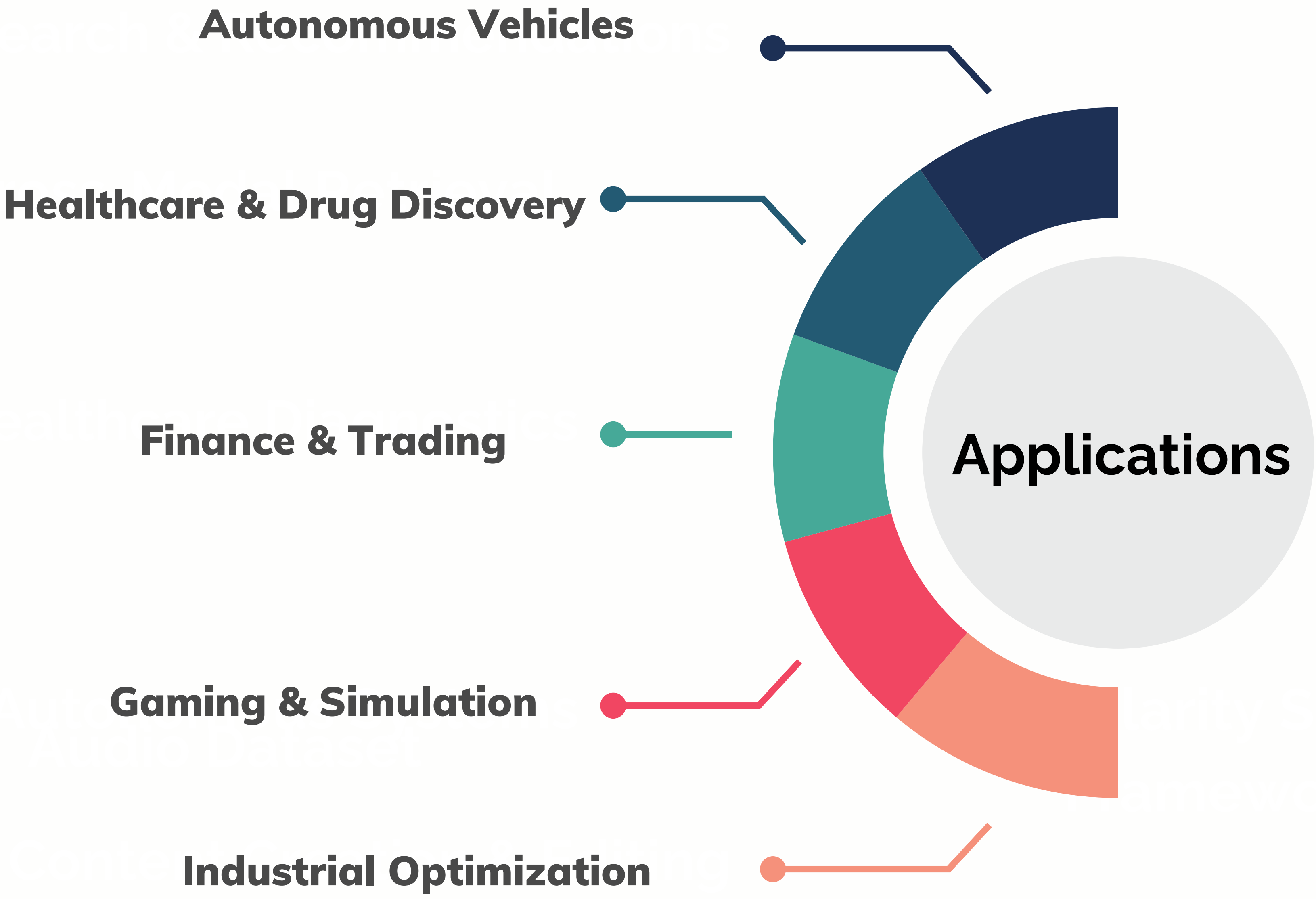
- **RL-Centric Training:**
  - GRPO treats verifiable reasoning steps (e.g., correct code/math solutions) as rewards. This drives the model to self-improve through iterative self-reflection.
  - Eliminates dependency on supervised fine-tuning (SFT) for initial training phases.
- **Hybrid Training Pipeline:**
  - Combines cold-start data (curated reasoning examples) with RL to stabilize learning and avoid poor readability issues.
  - Distills RL-trained 671B parameter models into smaller versions (1.5B–70B) while retaining 95%+ performance on tasks like code generation.
- **Performance Gains:**
  - Achieved 97.3% on MATH-500 (vs. OpenAI 01-1217's 96.4%) and 65.9% on LiveCodeBench coding tasks (vs. GPT-4o's 34.2%).
  - Outperforms non-RL models in faithfulness, with 59% of responses correctly vs. 7% for standard models.

BENEFITS


Improved Response Quality
01
02
Bias Reduction
Enhanced Adaptability
03
04
Optimized Long-Term Decision Making
Efficient Knowledge Distillation:
05

- **Improved Response Quality**: RL fine-tunes LLMs to generate more accurate and context-aware outputs.
- **Bias Reduction**: RLHF helps minimize biases by aligning models with diverse human feedback.
- **Enhanced Adaptability**: RL enables LLMs to refine responses dynamically based on user interactions.
- **Optimized Long-Term Decision Making**: RL helps LLMs prioritize meaningful responses over immediate rewards
- **Efficient Knowledge Distillation**: RL-trained models can be distilled into smaller, high-performing versions.

APPLICATION



Autonomous Vehicles

Healthcare & Drug Discovery

Finance & Trading

Applications

Gaming & Simulation

Industrial Optimization

- **Autonomous Vehicles:** Self-driving cars navigate traffic, avoid collisions, and optimize routes; AI controls traffic signals.
- **Healthcare & Drug Discovery:** AI designs treatment plans and discovers drugs by analyzing molecular interactions.
- **Finance & Trading:** RL optimizes high-frequency trading and dynamically balances investment portfolios.
- **Gaming & Simulation:** AI masters strategy games like AlphaGo and trains in simulations before real-world tasks.
- **Industrial Optimization:** Smart grids manage energy distribution, and AI enhances supply chain efficiency.

LIMITATIONS

BIAS REINFORCEMENT
HIGH COMPUTATIONAL COST
EXPLORATION VS. EXPLOITATION
REWARD DESIGN COMPLEXITY

- **Bias Reinforcement**: RLHF can amplify biases in training data, leading to unfair or skewed model outputs.
- **Exploration vs. Exploitation**: LLMs struggle to balance generating diverse responses (exploration) and refining known patterns (exploitation).
- **Reward Design Complexity** : Defining clear, effective reward signals for language tasks is challenging, often leading to suboptimal learning.
- **High Computational Cost:** RL-based fine-tuning demands extensive resources, making it expensive and less accessible.

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST