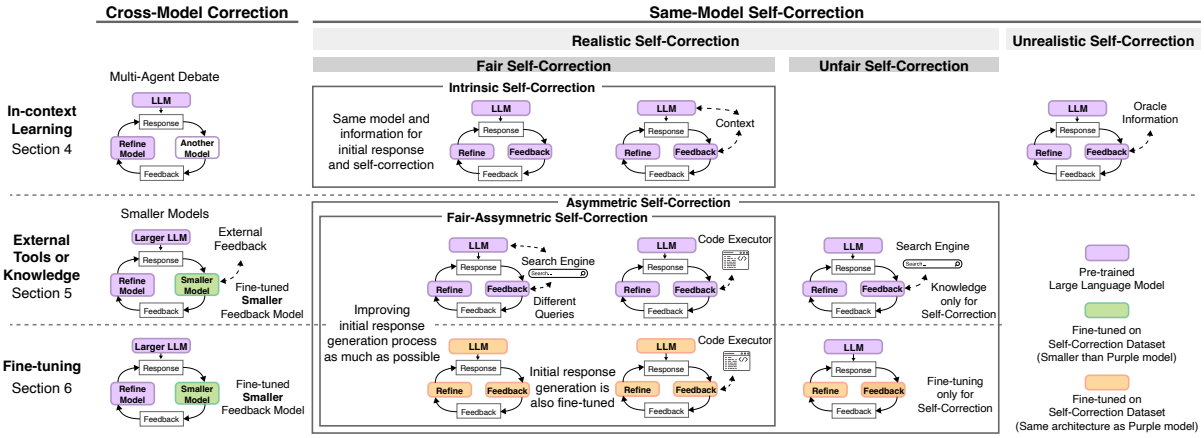

Figure 2: LLM self-correction frameworks, categorized by information used for generating feedback and whether they use best-possible initial responses (§3.2). This figure illustrates representative architectures.

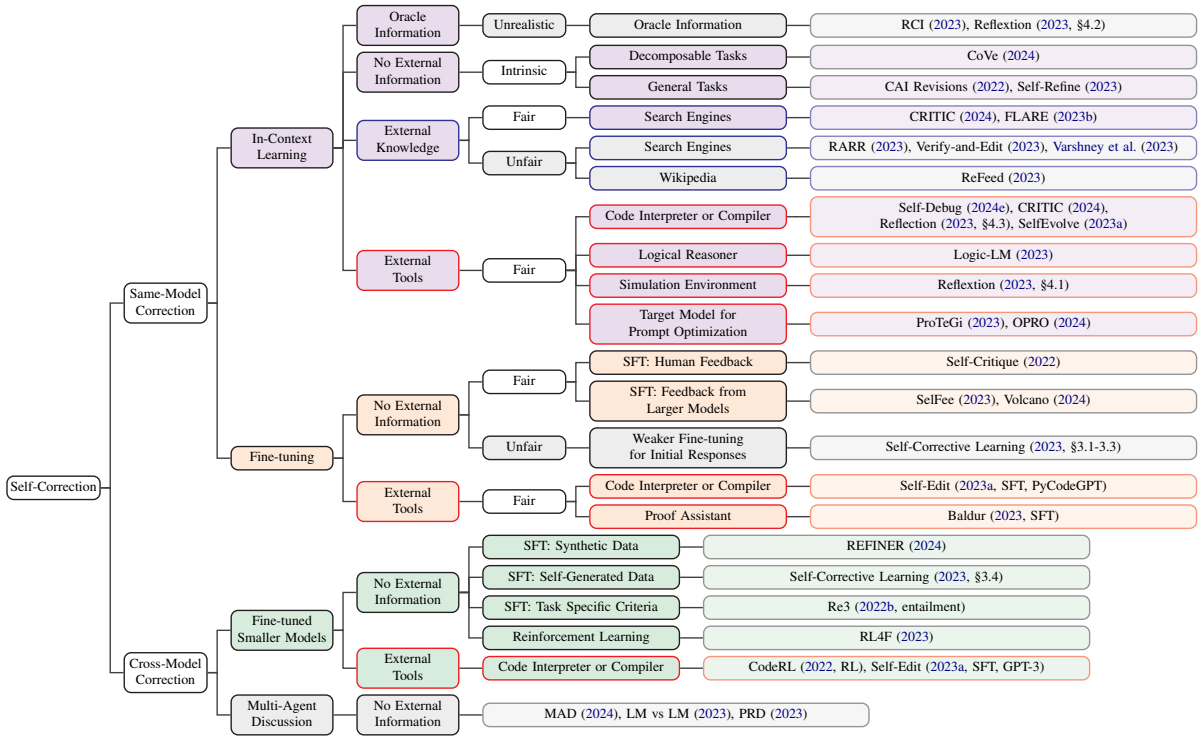

Figure 3: Taxonomy of LLM self-correction, categorized by information used for generating feedback and whether they use best-possible initial responses (fair or unfair). Refer to Section 3.2 for the definitions.

specific metrics (Xu et al., 2023), **external knowledge** from search engines (Jiang et al., 2023b; Gao et al., 2023; Zhao et al., 2023), Wikipedia (Yu et al., 2023; Zhao et al., 2023), or other corpora (Peng et al., 2023; Zhao et al., 2023), **oracle information** such as ground-truth answers (Kim et al., 2023; Shinn et al., 2023), human feedback (Chen et al., 2024a), or stronger models (Zhang et al., 2024c).

Fine-tuning (§5.2). Models fine-tuned for self-correction are another source of feedback, which are trained via supervised fine-tuning (Welleck et al., 2023; Ye et al., 2023; First et al., 2023; Paul et al., 2024; Han et al., 2024; Havrilla et al., 2024) or reinforcement learning (Le et al., 2022; Akyurek et al., 2023).

2.3 Tasks

Self-correction has been studied in various tasks, including **Reasoning:** arithmetic reasoning (Madaan et al., 2023; Nathani et al., 2023; Gou et al., 2024), code generation (Jiang et al., 2023a; Charalambous et al., 2023; Gou et al., 2024; Chen et al., 2024e;

RQ	Self-Refine (2023)	Huang et al. (2024a)	RCI (2023, §3.1)	RCI (2023, §3.2)	CRITIC (2024, §4.2)	CRITIC (2024, §4.3)	RARR (2023)
RQ1	✓	✗ (§3,5)	✓	–	✗	✗	–
RQ2	–	–	–	✓	✓	✓	–
RQ3	–	✗ (§4)	–	✓	–	✓	✓

Table 2: Research questions that prior studies implicitly target by claiming they are ✓ verified or ✗ refuted.

RQ	Requirements for Frameworks			Required Experiments	
	Information Symmetricity	Best-possible Initial Responses	Realistic	Comparison to Initial Responses	Comparison to Strong Baselines
RQ1	✓	✓	✓	✓	–
RQ2	–	✓	✓	✓	–
RQ3	–	–	✓	–	✓

Table 3: Requirements for experiments to verify each research question in Section 3.1.

Olausson et al., 2024), proof generation (First et al., 2023), logical reasoning (Pan et al., 2023), **Knowledge:** closed-book QA (Shinn et al., 2023; Gao et al., 2023; Jiang et al., 2023b; Gou et al., 2024), **Context-based Generation:** dialogue generation (Madaan et al., 2023; Peng et al., 2023), text summarization (Saunders et al., 2022), **Open-ended Generation:** conditional text generation (Ye et al., 2023; Schick et al., 2023), story generation (Yang et al., 2022b), detoxification (Schick et al., 2021; Bai et al., 2022; Gou et al., 2024; Phute et al., 2024), **Others:** machine translation (Chen et al., 2024c; Raunak et al., 2023; Ki and Carpuat, 2024), information retrieval (Gero et al., 2023), vision language tasks (Yin et al., 2023; Ge et al., 2023; Zhou et al., 2024; Lee et al., 2024; Huang et al., 2024b; Liu et al., 2024), and prompt optimization (Pryzant et al., 2023; Mehrabi et al., 2024; Yang et al., 2024).

2.4 Differences from Related Approaches

In this work, we define self-consistency (Wang et al., 2023) or generate-and-rank (Shen et al., 2021; Weng et al., 2023) to be different from self-correction because these approaches do not refine responses and assume that LLMs generate correct answers with a reasonable probability. We discuss these methods in Section 6 as strong baselines that should be compared with self-correction.

3 Research Questions

We find that prior studies often do not define their research questions in detail and fail to use appropriate self-correction frameworks in their experiments. We propose a new approach to classify research questions and frameworks in self-correction.

3.1 RQs in Self-Correction Research

Prior studies often simply state their research questions as *whether LLMs can self-correct their mistakes* (e.g., Kim et al., 2023; Madaan et al., 2023). However, we claim that research questions in self-correction research should be defined in more detail. We identify the following research questions implicitly targeted in prior studies, as in Table 2.

- [RQ1] Can LLMs self-correct their best-possible initial responses *based solely on the inherent capabilities?* (§4)
- [RQ2] Can LLMs self-correct their best-possible initial responses *assisted by external information?* (§5)
- [RQ3] Are the final outputs from self-correction *better than other methods?* (§6)

We define the *best-possible initial responses* as initial responses generated with best effort, using information that self-correction modules can access, such as external tools, knowledge, or fine-tuning.

Requirements for Verifying RQs. Experiments for verifying these research questions need to satisfy different requirements, as shown in Table 3. **External Information:** RQ1 needs to be evaluated on frameworks that refine responses using the same model without additional information. RQ2 and RQ3 can be evaluated on frameworks that use external information. **Initial Responses:** RQ1 and RQ2 need to be evaluated on frameworks that use the *best-possible initial responses*. RQ3 is about the final performance, so it is not necessary to start from strong initial responses. **Evaluation:** RQ1 and RQ2 only require to show that self-correction improves performance from the initial responses. RQ3 requires comparison with strong baselines (§6).

Confusion in Prior Work. Some prior studies implicitly target different research questions in a single work without clearly distinguishing them.

As in Table 2, Kim et al. (2023) target RQ1 for arithmetic reasoning by comparing self-corrected responses only with initial responses, but they target RQ3 for MiniWoB++ by comparing self-correction with baseline methods. Similarly, Gou et al. (2024) target RQ2 for arithmetic reasoning but target RQ3 for detoxification.

3.2 Frameworks for Verifying RQs

Prior work often categorizes self-correction frameworks based on approaches for generating feedback (§2). However, we point out that we also need to categorize them by the quality of initial responses because the frameworks we need to use for verifying different research questions vary by whether they use the best-possible initial responses (§3.1).

We propose categories of (same-model) self-correction that correspond to different research questions (§3.1), as shown in Figure 2. Specifically, we propose to categorize the self-correction frameworks as follows.

- Realistic: Can be used in real-world applications.
  - Fair: Using best-possible initial responses
  - Unfair: Using sub-optimal initial responses
- Unrealistic: Using information that is not accessible in real-world applications.

In this work, we focus on categorizing self-correction frameworks that do not involve multiple language models with different architectures. Cross-model correction uses different models for initial response generation and self-correction, so it is unsuitable for evaluating whether LLMs can improve their own initial responses [RQ1, RQ2]. However, it can be used to evaluate [RQ3] whether the final responses from self-correction are better than other methods.

Realistic vs. Unrealistic. Some prior studies propose unrealistic self-correction, which cannot be implemented in real-world applications, by using oracle information such as ground-truth answers (Kim et al., 2023; Shinn et al., 2023). These methods cannot be used to verify any research questions.

Fair vs. Unfair. Realistic frameworks can be categorized by whether they use the best-possible initial responses. **Fair self-correction** represents frameworks that refine the best-possible initial responses. (1) *Intrinsic self-correction* (Huang et al., 2024a) uses the same model and information for initial response generation and self-correction. Intrinsic self-correction can be used to assess [RQ1] whether LLMs can self-correct based solely on their inherent capabilities. (2) *Fair-asymmetric self-correction* uses additional information for self-correction, but also uses information to improve initial response generation as much as possible. For example, self-correction with code interpreters (Chen et al., 2024e; Gou et al., 2024) is not intrinsic but fair because we cannot easily use code interpreters to directly improve the initial response generation. Fair-asymmetric self-correction can be used to evaluate [RQ2] whether LLMs can self-correct the best-possible initial responses using external information. **Unfair self-correction** (or *unfair-asymmetric self-correction*) represents frameworks that are practical but do not use the best-possible initial responses. For example, methods that use search engines only for self-correction (Gao et al., 2023; Yu et al., 2023) are unfair because they can use search engines to directly improve the initial response generation. Unfair self-correction can evaluate [RQ3] whether the final responses from self-correction outperform other methods but cannot evaluate [RQ2] whether self-correction can improve the best-possible initial responses.

4 Self-Correction with Prompting

> [RQ1] Can LLMs self-correct their best-possible initial responses *based solely on the inherent capabilities?*

Several studies propose *intrinsic self-correction* methods, which self-correct responses from LLMs by prompting themselves to generate feedback and refine the responses. Bai et al. (2022) propose self-correcting harmful responses from LLMs by prompting themselves. Self-Refine (Madaan et al., 2023) and RCI Prompting (Kim et al., 2023) iteratively prompt LLMs to self-correct their own responses in tasks such as arithmetic reasoning.

Negative Results. However, recent studies report that intrinsic self-correction does not improve or even degrade the performance in tasks such as arithmetic reasoning, closed-book QA (Huang et al., 2024a; Gou et al., 2024), code generation (Gou et al., 2024; Olausson et al., 2024), plan generation (Valmeekam et al., 2023), and graph coloring (Stechly et al., 2023). Several studies claim that a bottleneck is in the feedback generation, and it is