Let Me Teach You: Pedagogical Foundations of Feedback for Language Models

Beatriz Borges[1], **Niket Tandon**[2], **Tanja Käser**[1], and **Antoine Bosselut**[1]

[1]EPFL [2]Allen Institute for Artificial Intelligence
{beatriz.borges, antoine.bosselut}@epfl.ch

Abstract

Natural Language Feedback (NLF) is an increasingly popular avenue to align Large Language Models (LLMs) to human preferences. Despite the richness and diversity of the information it can convey, NLF is often hand-designed and arbitrary. In a different world, research in pedagogy has long established several effective feedback models. In this opinion piece, we compile ideas from pedagogy to introduce FELT, a feedback framework for LLMs that outlines the various characteristics of the feedback space, and a feedback content taxonomy based on these variables. Our taxonomy offers both a general mapping of the feedback space, as well as pedagogy-established discrete categories, allowing us to empirically demonstrate the impact of different feedback types on revised generations. In addition to streamlining existing NLF designs, FELT also brings out new, unexplored directions for research in NLF. We make our taxonomy available to the community, providing guides and examples for mapping our categorizations to future resources.

1 Introduction

The last few years introduced a new paradigm fine-tuning training Large Language Models (LLMs): training them with human feedback (Ziegler et al., 2020; Stiennon et al., 2022; Bai et al., 2022a; OpenAI, 2023) to augment their capabilities what they learned during pretraining (Christiano et al., 2017; Wu et al., 2021; Menick et al., 2022). Pursuing this alignment has led to models that (1) are less toxic and harmful (Bai et al., 2022b; Korbak et al., 2023), (2) have the ability to morally self-correct (Ganguli et al., 2023), and (3) are preferred by their users (Ouyang et al., 2022a; Bai et al., 2022b).

The currently most popular approach for aligning models with feedback is Reinforcement Learning from Feedback (RLF). Be it using human feedback (RLHF; Ouyang et al., 2022a; Bai et al., 2022a; Ramamurthy et al., 2023) or feedback generated by LLMs (RLAIF; Bai et al., 2022b; Saunders et al., 2022; Madaan et al., 2023), both have been shown to be successful in several metrics — from encouraging honest behaviors, to reducing toxicity, to being overall preferred by evaluators (Ouyang et al., 2022a). Other approaches, such as imitation learning (Li et al., 2016; Kreutzer et al., 2018; Hancock et al., 2019; Scheurer et al., 2022), and feedback modelling (Weston, 2016; Li et al., 2017; Hancock et al., 2019; Xu et al., 2022; Liu et al., 2023) have had similar success. Feedback has thus emerged as an important source of information for

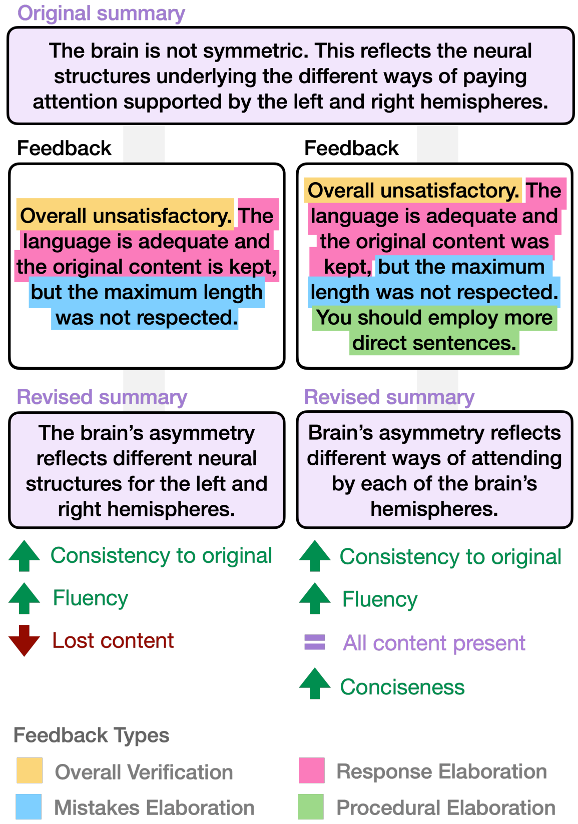

Figure 1: An example task with the provided feedback classified using categories from our taxonomy (§5). Different types of feedback produce different responses, motivating the systematic study of feedback types in natural language feedback research.

model improvement and evaluation, ensuring that models progress toward desired objectives and behaviors (Fernandes et al., 2023).

Despite the observed benefits of learning from feedback, a systematic study of what constitutes helpful feedback remains absent from the ongoing discussion on this topic. For example, RLF requires a Reward Model (RM) to be trained on numerical or ranking-based feedback data (Rafailov et al., 2023) — a format that is limited in the amount of information it conveys (Wu et al., 2023). To counteract this limitation, works have begun exploring Natural Language Feedback (NLF; Weston, 2016; Madaan et al., 2023; Wu et al., 2023). However, these works rely on the "intuitive guesses" of authors regarding what constitutes useful feedback. This approach leads different works to explore distinct conceptualizations of NLF, preventing a systematic comparison. Given its importance in improving LLMs, the lack of systematicity in feedback design and analysis represents a current blind spot in NLP.

Meanwhile, in the domain of learning sciences, there exist comprehensive studies of various aspects of feedback, as feedback is considered an essential component of instruction and learning. In this opinion piece, we propose to ground the study of LLM feedback to pedagogy research, and to adapt this foundation for LLMs. To this end, we first present the most relevant feedback-related models from the learning sciences (§3). Taking inspiration from these pedagogical models, we create a novel framework, FELT, that expansively maps the various features of the feedback space, incorporating dimensions exclusive to LLMs, such as prompt information and model characteristics (§4).

We then introduce a comprehensive, grounded taxonomy of feedback content, with 10 distinct dimensions, defining several important aspects of feedback that remain underexplored (§5.1). To make our taxonomy accessible, we further propose an interpretable categorization of feedback types, specifically designed to be modular and applicable to future works on feedback for LLMs (§5.2). We use our dual taxonomy to classify feedback types in prior studies on LLM feedback NLP, as well as to study their effect in a case study on an open-ended summarization task (§6). The application of our taxonomy shines a light on the underspecification of current approaches to feedback formulation, and suggests promising areas of future research.

Contributions We summarize our contributions as follows: (i) We present a survey of pedagogical feedback formulations and models; (ii) we outline the variables that influence feedback and its processing into a schematic framework, specifically adapted for LLMs; (iii) we propose a general and extensive taxonomy of feedback content, as well as a more condensed collection of emergent feedback content categories; (iv) we show that different feedback types produce distinctive changes in the revised generations of LLMs; and (v) we map previous research to our taxonomy, and provide guides and examples for stratification of feedback analysis using our categorization, thereby providing tools for future investigation of how feedback types influence LLM performance.

2 Feedback in NLP

2.1 How informative is feedback for LLMs?

The value of feedback is derived from the implicit information it represents about human values and expectations, that would otherwise be extremely difficult to specify (Christiano et al., 2017). While all forms of feedback are able to reflect this knowledge to some degree, not all of them can represent the same amount and granularity of information.

Feedback Representation Feedback can assume different forms: numerical ratings, rankings, preferences, demonstrations, and fully textual information (which can either be based on a rigid template or unconstrained, free-form text — structured and unstructured feedback, respectively).

Learning From Feedback The most popular RLHF methodologies usually collect either a numeric rating or ranking from human workers for classifying the *quality* of feedback (typically focused on encouraging *helpfulness* and *honesty* while mitigating *harmfulness*; Askell et al., 2021). RLF may also leverage demonstrations to finetune LLMs in a supervised fashion before the RLF stage takes place to reduce the subsequent search space (Ouyang et al., 2022a; Bai et al., 2022b), with scalar or ranking data subsequently used to train the reward model. This approach attempts to address the intractable problem of designing an appropriate loss function to express the aforementioned goal of honest, helpful and harmless language models (Askell et al., 2021).

Feedback Alignment However, the extent to which feedback (i.e.,, information that LLMs are

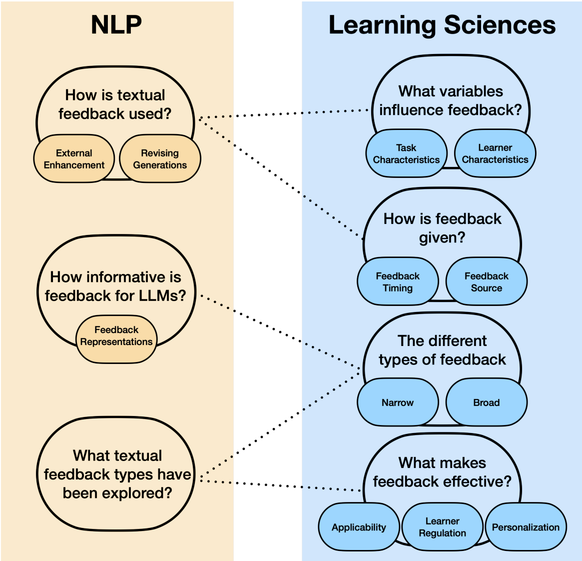

Figure 2: Connecting feedback research in NLP to foundations of feedback in the Learning Sciences.

finetuned on) transmits these goals remains unclear. For example, InstructGPT[1] (Ouyang et al., 2022b) is finetuned on demonstration data, and subsequently trained using RLHF with a reward trained using comparison data (i.e.,, specifically, pairs of ranked generations). This feedback is limited in the amount of information it transmits. For a given prompt, marking demonstration A as better[2] than demonstration B provides little information on the quality of A nor B, nor on whether A fully outclasses B, or whether B may surpass A in some some dimensions. In any case, such a format provides no information on how either demonstration can be improved. Taking both these limitations and human bias into account, RMs are likely to suffer from some degree of distortion and misalignment. Other approaches (Liu et al., 2023; Gao et al., 2022) to model training with human feedback also still rely on simple ranking or numerical feedback. Constitutional AI (Bai et al., 2022b) employs a similar approach, but the feedback — both the textual feedback used for initial supervised finetuning and the ranking feedback used to train the RM — is generated by LLMs[3] rather than human workers (i.e.,, RLAIF). While using LLM-generated demonstrations makes the method more scalable for data collection, the same challenges of remain.

[1]And later OpenAI models such as `text-davinci-003`.

[2]We note such a format also obfuscates any bias and disagreement that occurred in reaching such a judgment

[3]Only the harmlessness feedback is generated by an LLM, human feedback is used for the helpfulness dimension.

Informational Limitations Recent works have started to acknowledge the limited information in the aforementioned feedback formulations, recognizing them as unsuited for capturing critically relevant information, such as different types of errors (Golovneva et al., 2023; Wu et al., 2023).

2.2 How can textual feedback improve LLMs?

The most commonly used feedback formulations, scalar and ranking feedback, are thus limited in the information they can convey. An intuitive alternative is to instead leverage textual feedback.

External Enhancement Augmenting the model externally — be it through data augmentation (Shi et al., 2022), external corrective feedback (Tandon et al., 2022; Madaan et al., 2022; Shinn et al., 2023) or natural language patches (Murty et al., 2022) — is one relatively straightforward approach to incorporating textual feedback into a LLM.

Revising Generations Various works have instead introduced a secondary model, that either refines an original LLM's answer (Scheurer et al., 2022; Welleck et al., 2022; Tandon et al., 2022), critiques it (Saunders et al., 2022; Paul et al., 2023) or iteratively self-improves (Schick et al., 2022; Chen et al., 2023; Madaan et al., 2023). Several of these approaches leverage the same LLM for both the original answer generation as well as its refinement, but all of them rely on textual feedback — be it for eventual dataset augmentation and refinement, or as part of the input to the new answer revision. All these models, beyond leveraging some kind of natural language feedback, also target intermediate generations with their feedback, not the final outcome. This intermediate feedback is another mechanism to transmit more information to a model. Rather than increasing the feedback complexity, these approaches increase the number of feedback opportunities, through multiple iterations (Lightman et al., 2023).

What is missing? A clear trend towards more informative feedback is underway, drifting away from the still dominant approach of reducing feedback to a single scalar or ranking. However, the textual feedback employed by different works are often completely different from one another. No work so far has taken up a true mapping of the feedback space, identified the different types of information that can be encoded in NLF, and allowed for an exploration of different feedback components and their effectiveness.

2.3 What types of textual feedback have been explored in NLP?

Given the poverty of feedback forms used to train LLMs, a variety of works have recently emerged to use natural language feedback to correct LLMs, but this area remains in its infancy. A recent survey on how feedback is receive and integrated with LLMs (Fernandes et al., 2023) recognizes the limitation of current score-based approaches to feedback, and proposes that future work should leverage the much richer signal of NLF. Shi et al. (2022) distinguishes textual feedback depending on whether the feedback is being formally provided for the model's answer, or whether, remaining in the dialog setting, the user mentions they disliked the reply they received. SELF-REFINE (Madaan et al., 2023) argues that the quality of the generated feedback is critical, though they only compare their "actionable and specific" LLM-generated feedback against "generic feedback" and the complete absence of feedback in an ablation study. Wu et al. (2023) propose the introduction of finer-grained feedback at sub-sentence, sentence and full sentence levels — and of three different error types: factual incorrectness, irrelevance, and information incompleteness. Despite the impressive performance of this approach, the feedback exploration is limited at only three specific types, and only preference rankings are used. Finally, Weston (2016) conducted the most thorough exploration to date, exploring 10 different dialogue-based supervision modes, which represent different interaction and feedback types. However, these modes often overlap information-wise, limiting the conclusions of the study.

3 Feedback in Education

In this paper, we study feedback in human learning to construct a comprehensive, theory-grounded feedback taxonomy that directly addresses the limited exploration of natural language feedback. We build off the work of Lipnevich and Panadero (2021), who conducted a systematic review of work in the fields of education, psychology, information processing and assessment philosophy, to eventually select the 15 most relevant and influential works on feedback models research. In this section, we provide a brief overview of the key points of each of these works — related to the definition, effectiveness, and characteristics of feedback — and draw inspiration from them to subsequently propose a framework for feedback integration (§4) and a taxonomy for feedback content (§5).

3.1 What is feedback?

Many prior works have proposed a definition of feedback, and all agree that feedback either is information or contains information provided to a learner.[4] Consensus starts to wane on the other properties feedback must possess, one of which we note as particularly interesting: roughly half of the feedback models (Ramaprasad, 1983; Butler and Winne, 1995; Narciss and Huth, 2004; Narciss, 2008; Nicol and Macfarlane-Dick, 2006; Hattie and Timperley, 2007; Lipnevich and Panadero, 2021; Panadero and Lipnevich, 2022) incorporate the idea of a *gap*, stating that feedback should provide the learner with information about the difference between their actual performance and the target performance. In contrast, the remaining models do not explicitly bridge the necessity of a performance gap in their formulation of feedback.

Defining Feedback As a result, while different studies may disagree on the breadth or specificity required for feedback, and the limitations on its content, purpose or effect to be considered feedback, a definition (which we adopt throughout this paper) nevertheless emerges from their points of consensus:[5] *any task-relevant information given to a learner, by any possible feedback-generating agent (including internal feedback)*. Note that we do not impose any constraints on the information that is given to the learner.

3.2 What constitutes effective feedback?

Kluger and DeNisi (1996) showed that in 38% analyzed cases, feedback had a detrimental effect on a learner's performance, challenging the intuitive understanding of feedback as helpful information. This reality requires reflecting on the properties of feedback that effectively help learners improve. Three main conditions for ensuring helpful feedback have emerged from previous work: *applicability*, *learner regulation*, and *personalization*.

[4] However, this definition is not a sufficient condition for some of these works. Carless and Boud (2018), for example, reflects a learner-centric perspective, viewing feedback as the process through which the student understands and integrates information — thus, without the student processing, there is no feedback even if the information is present.

[5] For an overview of all the different definitions of feedback discussed, please see Appendix A.

Applicability Feedback should be actionable, i.e., should support the learner in achieving the target performance. Applicable feedback develops naturally from the directives about clarifying the task's objective, providing quality information, and creating opportunities for the learner to improve. Sadler (1989) therefore suggest that feedback needs to identify a target performance, compare the learner's current performance to it, and engage in actions to reduce that difference. Similarly, Hattie and Timperley (2007) indicate that effective feedback needs to answer three questions: where the learner is going (the goal), how they can get there, and where to go next. Other works extend these definitions of effective feedback by including elements such as motivational and metacognitive aspects (Nicol and Macfarlane-Dick, 2006) or aspects of teaching (e.g., lesson design; Evans, 2013).

Learner Regulation Effective feedback produces a positive response in the learner. Kluger and DeNisi (1996) argue that, in response to feedback, a learner's attention will be directed to one of three levels: how to solve the task, the task as a whole, or meta-task processes (processes the learner is doing while performing the task). Others (Nicol and Macfarlane-Dick, 2006; Evans, 2013) further note that effective feedback also enhances self-regulated learning behaviors. Narciss and Huth (2004); Narciss (2008) extend this definition by arguing that feedback can have three distinct types of impact: influence on the learner's cognitive abilities, their metacognitive skills, or their motivation and self-regulation. Lipnevich et al. (2016) defend that when a student receives feedback, they produce cognitive and affective responses. Namely, the learner will judge how worthwhile the task is, how much they control the outcome, and how understandable the feedback is. In turn, this judgment produces a behavioral reaction, influencing their performance and learning. Similarly, Panadero and Lipnevich (2022) state that feedback impacts both the students' performance and learning as well as their affective processes and self-regulation.

Personalization Different types of feedback are best suited for different learner characteristics and should be adapted accordingly (Mason and Bruning, 2001). Furthermore, the learner's individual characteristics will directly impact how the process feedback (Lipnevich et al., 2016).

We conclude that not all feedback is good feedback, though different models propose different rule sets for achieving effective feedback, suggesting a need for further exploration. Furthermore, feedback can be given with different purposes, from directly improving the learner's performance, to clarifying the task, to improving metacognitive skills, and to help them regulate their emotions, motivation, and inner-processes. We revisit this concept of feedback purpose in Section 5.

3.3 What are the characteristics of feedback?

In Section 3.2, we summarized different properties of feedback, observing that not all types of feedback are effective for learning in every situation. A large body of research has therefore attempted to systematically categorize feedback based on its *content*, how it is given (*timing*, *source*), and the variables influencing it (*task*, *learner*).

3.3.1 What are the types of feedback?

While there is a plethora of work on systematically categorizing feedback, previous work can be broadly divided into two groups: taxonomies of feedback focusing on the content of the feedback only and taxonomies taking into account the whole ecosystem of the feedback.[6]

Narrow Works in this category focus on the characteristics of the content only. Kulhavy and Stock (1989), for example, model feedback through a verification component, which is a simple discrete classification of the answer as correct or incorrect, and an elaboration component, which contains all other information. Other works (Hattie and Timperley, 2007; Panadero and Lipnevich, 2022) suggest three categories for classifying feedback: (i) addressing the learner's performance goal, (ii) addressing the learner's current performance, and (iii) addressing the next steps the learner should undertake.

Broad In contrast to the first group, works in the second group suggest a more comprehensive categorization of feedback including the whole feedback ecosystem. For example, several works propose different feedback categories that take into account characteristics of the learner (student proficiency, prior knowledge) and the task (difficulty) (Mason and Bruning, 2001; Narciss and Huth, 2004; Narciss, 2008).

[6] Appendix A presents a more thorough definition of proposed feedback categories in the learning sciences.

3.3.2 How is feedback given?

Apart from its content and the ecosystem around it, feedback has also been characterized by the manner in which it is given. Two main components have emerged in the literature.

Source Feedback can be given by different sources (e.g.,, teachers, peers, or even the learner themself). In their systematic review, Lipnevich and Panadero (2021) found that seven out of the 15 considered models view the source as an important characteristic of feedback. An additional three works distinguish feedback generated by an external source from feedback generated internally.

Timing The timing of feedback is considered essential too, with previous differentiating between immediate feedback and delayed feedback. While early work (Bangert-Drowns et al., 1991) found delayed feedback to be more effective, more recent works (Mason and Bruning, 2001) argue that the optimal timing of feedback depends on learner characteristics. For example, Hattie and Timperley (2007) state that the most beneficial timing depends mainly on the complexity of the task; complex tasks benefit from delayed feedback as they allow the learner to properly process the task. Other works (Narciss, 2008) focus mainly on the learner characteristics, arguing that as long as the learner possesses the metacognitive skills required to spot and address mistakes, feedback should be delayed.

We incorporate both feedback source and timing in our framework presented in Section 4. Other points of contention remain beyond these two dimensions, such as feedback valence, but consider valence as an element of the feedback's content, and as such discuss it only in Section 5.

3.3.3 What variables influence feedback?

Finally, feedback cannot only be categorized according by its content and the way it is given, but is also characterized by the ecosystem surrounding it. In the works cited in the previous two sections, many authors mention the challenge of determining optimal feedback type in isolation. Instead, the characteristics of the *task* and the *learner* are important to take into account when giving feedback.

Task The characteristics of a task have been shown to influence the optimal timing of feedback (see Section 3.3.2) as well as the content of the feedback. In particular, Mason and Bruning (2001) takes into account the complexity of the task when choosing the most suitable feedback for a given setting and Narciss and Huth (2004); Narciss (2008) incorporate the task and the instructional content and goals into the instructional factors that affect feedback. Also the nature of the task (e.g., closed versus open-answer) influences feedback content and processing (Lipnevich and Panadero, 2021).

Learner Previous work has also acknowledged the impact of learner characteristics on the effectiveness of feedback. Mason and Bruning (2001), for example, consider student achievement and prior knowledge as important factors directly impacting the best suited type of feedback. Nicol and Macfarlane-Dick (2006) expands the learner's prior knowledge and proficiency into *domain knowledge* and *strategy knowledge* (along with *motivational beliefs*), which are updated upon each attempt through both internal and external feedback.[7] Narciss and Huth (2004) and Narciss (2008) flesh out the learner characteristics even further, including factors such as learning goals and motivation. Similarly, Lipnevich et al. (2016) identify "personality, general cognitive ability, receptivity to feedback, prior knowledge, and motivation" as key learner characteristics that impact feedback processing.

Other works directly focus on feedback processing mechanisms of the learner. Kluger and DeNisi (1996) propose three processing levels: details about how to solve the task, the task itself as whole, and meta-task processes. In contrast, Hattie and Timperley (2007) present four levels of feedback processing: (1) *task level*, conveying how well the tasks are understood and performed, (2) *process level*, the process needed to achieve the *task level* understanding and performance, (3) *self-regulation level*, related to self-monitoring and the direction and regulation of the learner's actions, and (4) *self level*, which reflects personal evaluations about the learner as a whole.

4 Unifying The Two Worlds

Feedback emerges as both a complex ecosystem, and a rich, but systematized, information source, with many attributes covered in educational work. We consolidate several of these features, introduced in Section 3, to create a novel feedback framework, which we subsequently adapt for LLMs.

[7] Information for internal feedback generation is derived not only from the learner's initial state but also from their goals, tactics and strategies and their (internal) learning outcomes — through self-regulatory processes.

4.1 The Unified Framework

From the multi-disciplinary background presented in Section 3, we derive a tripartite feedback ecosystem structure consisting of task, learner, and feedback components, depicted in Figure 3.

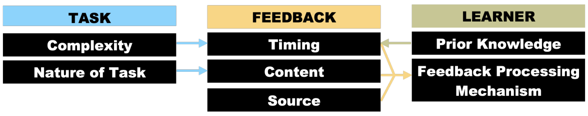

Figure 3: The feedback ecosystem. The feedback's characteristics, the task, and the learner all influence how effectively the feedback is received.

Learning Sciences Grounding Each component has a set of attributes that can vary depending on a situation requiring feedback. As discussed in Section 3, feedback is influenced by both *task* and *learner characteristics*. We break each of these two components down to more granular constituents. For the *task* we introduce two attributes, its complexity (i.e., its difficulty level), and its nature — which we reduce to it being closed-answer (where there is a single correct answer or a finite set of them) or open-ended. The *learner* is similarly divided into two sub-components: their particular feedback processing mechanism, and their prior knowledge (dependent on the task). We also develop a more holistic view of the *feedback* component itself, with three main features:

(i) *timing*, whether feedback is provided immediately or after a given temporal delay,

(ii) *content*, capturing both the type and the format of the information provided in a feedback message, explored fully in section 5, and,

(iii) *source*, whether the feedback stems from a peer or an authority figure, a human or an AI model, possibly including their relevant proficiency levels. Different sources usually build different relationships with the learner, and will communicate in different language *styles*, a dimension explored in Section 5.

Framework Interactions As Figure 3 depicts, various interactions occur between the task, the learner, and the feedback. For example, as stated in Section 3, Mason and Bruning (2001) argue that if a learner has little prior knowledge or the task has a low complexity, feedback should be immediate, but if the task complexity and student's prior knowledge are both high, then it should be delayed. Narciss (2008) on the other hand, refers to Mathan and Koedinger (2005)'s take on timing, which states that if the learner possesses metacognitive skills (for identifying and correcting errors), then feedback should be delayed, to first promote these skills. The type of task also naturally conditions the feedback given. Finally, all aspects of feedback impact how the learner receives and processes the feedback. Appendix B presents a more comprehensive overview of the interactions present in the framework.

4.2 The LLM Adaption – FELT

The unifying framework is directly derived from learning sciences research, and as such, designed for the human learner. We are, however, interested in providing feedback to LLMs. Thus, we adapt this ecosystem to LLMs, where the LLM is viewed as the learner. In particular, we propose five modifications to the base framework: expanding the learner to reflect the model's size as well as the model training data and method (including its training objectives), the expansion of the task component to reflect the prompt instructions, and finally, the addition of an error component. The resulting framework, FELT (Feedback, Errors, Learner, Task) is displayed in Figure 4.

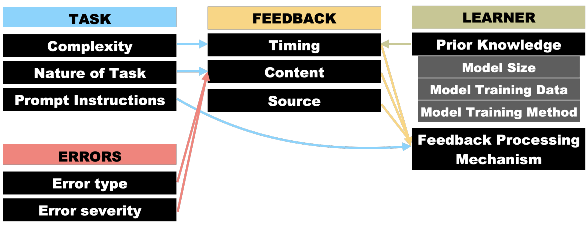

Figure 4: FELT framework. The feedback ecosystem is specifically adapted for LLMs.

Expanding the Learner We add three specific LLM extensions to the generic learner component: model size, and model training data and method. All three are important, and together capture the *Prior Knowledge* of a model. Model size is directly linked to emergent abilities (Wei et al., 2022) and the model's ability to effectively leverage feedback (Scheurer et al., 2022; Bai et al., 2022b). The data the model was trained on, as well as how it was trained, similarly encode the model's initial abilities to tackle any given task.

Expanding the Task Another important factor to incorporate into the framework pertains to the instructions given in the prompt. Past research has shown the importance of stating the actions a model can take, such as outputting "I don't know." (Zhou et al., 2023). Similarly, how strongly the prompt encourages a model to incorporate feedback can favor overoptimization.

Introducing Errors Finally, effective feedback may communicate information on where the learner is failing, requiring an understanding of the possible error modes for a given task, and which ones the learner is likely in. For example, guessing and committing systematic reasoning mistakes are reflections of differing understandings. Exploring the error space and identifying the mistakes made by a learner is an important extension to the base framework directly derived from pedagogical and psychology of education research.

4.3 Feedback Integration

The method used to transmit the feedback to the model influences how it is subsequently processed. Fernandes et al. (2023) identify three common feedback integration mechanisms: feedback-based imitation learning, joint-feedback modeling, and reinforcement learning. In addition to this, we also consider feedback use in in-context learning (Brown et al., 2020). The training objective will necessarily influence how the model is processing and incorporating feedback. Typically, the training relies upon either scalar feedback (a single number encoding how much the model should be rewarded for its output) or a ranking (how well a given output did in relation to other candidate answers). However, this is simple information, and does not leverage the rich and complex information encoded in natural language feedback. Section 5 therefore comprehensively explores the different types of information that can be encoded in feedback.

5 Feedback Content Taxonomy

In Section 4, we presented an overview of the complex ecosystem of feedback, including an expansion specifically for LLMs (i.e., FELT) that connects various background elements (e.g., the learner, the task, the error types) to the actual feedback that must be given. In this section, we expand on our analysis of the *content* dimension of feedback in FELT. Specifically, we present a taxonomy of feedback content under two different forms: a set of 10 broad axes along which feedback can vary, and a more concrete set of nine emergent categories for feedback topic. Figure 4 presents an overview of the two different presentations of this taxonomy, and the mapping between them.

We motivate this taxonomy to finely categorize current approaches to textual feedback that implicitly formulate feedback solely for *utility* (i.e., how useful is the feedback for guiding a model toward a suitable response). However, they do not categorize its content, leaving a conceptual gap about *what* makes feedback useful. Our taxonomy stratifies the feedback space, allowing a deliberate and systematic study of feedback content.

5.1 General Taxonomy

We break down feedback content along ten dimensions that influence how feedback is formulated:

1. *length*, an indication of how much feedback feedback is given, possibly measured by counting its number of tokens,
2. *granularity*, a measure of the level of detail with which the feedback addresses the original answer — it is not a measure of how much of the answer is being considered, but rather of the level of detail with which it is being considered,[8]
3. *applicability of instructions*, expressing both whether the feedback contains instructions, as well as how applicable those instructions are for the learner and their current understanding and approach to solving the task,
4. *answer coverage*, which registers how much of the learner's answer is considered to generate the given feedback. The feedback could be independent of the answer, or only relate to parts of the answer (e.g.,, focusing on a particular mistake), or the feedback might take the complete answer into consideration,
5. *criteria*, denoting which criteria the answer is being evaluated on: global evaluation, specific dimensions (e.g., fluency, engagement, etc.), or, alternatively, no dimensions (the answer is not being evaluated),
6. *information novelty*, indicating the degree to which learner already had access to the information provided in the feedback, ranging from all information being previously known

[8] For an open-answer example task, feedback might range from global learning meta-feedback, to global but task-specific, to paragraph-level, to sentence-level, to word-level, to token-level feedback.

	Length	Granularity	Applicability of Instructions	Answer Coverage	Criteria	Information Novelty	Purpose	Style	Valence	Mode
Global Verification	Short	Very Low	None	Full	Global	Partially Novel to Novel	Improve performance	All allowed	Positive OR Negative	All allowed
Response Verification	Short	Low	None	Full	Global to Specific	Partially Novel to Novel	Improve performance	All allowed	Positive OR Negative	All allowed
Mistakes Verification	Short	Low	None	Partial	Global to Specific	Partially Novel to Novel	Improve performance	All allowed	Negative	All allowed
Correct Answer	Short to Long	Very Low OR Very High	None	None OR Full	None	Partially Novel to Novel	Improve performance	All allowed	Neutral	All allowed
Response Elaboration	Medium to Long	Medium to High	None to Low	Full	Global to Specific	Partially Novel to Novel	Improve performance	All allowed	Positive, Negative, OR Both	All allowed
Mistakes Elaboration	Medium to Long	Medium to High	None to Low	None to Full	Global to Specific	Partially Novel to Novel	Improve performance	All allowed	Negative	All allowed
Task Elaboration	Short to Long	Very Low to Medium	None to Low	None to Full	None	Not Novel to Partially Novel	Clarify task	All allowed	Neutral	All allowed
Procedural Elaboration	Short to Long	Very Low to Medium	Low to High	None to Full	Global to Specific	Not Novel to Partially Novel	Improve performance	All allowed	Neutral-Negative	All allowed
Metacognition Elaboration	Short to Long	Very Low	None to Low	None	None	Not Novel to Partially Novel	Learn how to learn and act	All allowed	Neutral	All allowed

Figure 5: A mapping between the ten axes of our general taxonomy and the nine feedback content categories.

by the learner, to some information being unknown to the learner, to all information being novel for the learner,

7. *purpose*, measuring whether the feedback is being given to improve the learner's performance or to clarify the task,[9]

8. *style*, capturing the level of the language used to transmit the feedback to the learner, which can range from simple, direct sentences to verbose and terminology-heavy language,

9. *valence*, indicating whether the feedback is positive (signaling achievement) or negative (signaling need for improvement),

10. *mode*, denoting how the feedback is given to the learner, capturing two important aspects: whether the feedback is uni- or multi-modal, [10] and whether it allows the user to submit multiple tries or not.

Combined, these ten axes capture what we consider the most important qualities of a piece feedback to understand its impact on a given learner model. We hypothesize that each of these dimensions influences the model's revised response to varying degrees, but that all are worthy of individual study.

5.2 Categorical Taxonomy

Given these ten axes of feedback, we further define nine emergent, interpretable feedback categories that vary each of these axes (see Figure 5). These nine categories enable a more streamlined classification of prevalent forms of feedback, yielding a starting point to explore which components of feedback may be effective for a given task:

1. *Global Verification*, a single, aggregate score for the task performance as a whole,
2. *Response Verification*, response-level classification of the answer (e.g., *right* or *wrong)*,
3. *Mistakes Verification*, error-level feedback, possibly with the number of mistakes,
4. *Correct Answer*, provides the correct answer,[11]
5. *Response Elaboration*, response-level, response-specific feedback addressing either the positive characteristics of the answer, its shortcomings, or both,
6. *Mistakes Elaboration*, elaboration feedback about types and sources of errors,
7. *Task Elaboration*, elaboration feedback about the task, such as its requirements, topic, the relevant concepts or terminology — this is clarification feedback, and does not provide suggestions for the next steps,
8. *Procedural Elaboration*, task-specific elaboration feedback about how to solve or improve a solution for said task,

[9] In a pedagogical setting with human learners, other purposes are possible, such as regulating the student's emotions and motivation, but we do not consider these for the LLMs.

[10] Multi-modal feedback is naturally more suited for multimodal tasks. For example, in an instance segmentation task, the correct (visual) answer could be provided alongside textual feedback on mistakes and how to correct them.

[11] While understanding the concept of the correct answer in a closed-answer task is trivial, the same cannot be said about an open-answer context. Providing an excellent answer for the task would be nothing more than a worked example and thus a *Procedural Elaboration* type of feedback. Instead, in this context, the *Correct Answer* becomes the rewritten version of the student's answer, improved to be excellent according to the evaluative standard, ideally with as few changes as possible.

9. *Metacognition Elaboration*, elaboration feedback about general learning and problem-solving strategies

Examples of each feedback type, as well as more detailed guidance on the characteristics of each type, are presented in Appendix C. The nine categories, with the exception of the *correct answer*, are grouped into two main types: verification and elaboration. Following the terminology established by educational works, verification feedback provides classification-only information, such as whether an answer is right or wrong, or whether a mistake was made or not. Elaboration feedback provides more meaningful, concrete, and detailed information.

In our formulation, the general taxonomy (§5.1) provides a set of 10 broad feedback dimensions that the categorical taxonomy composes into nine distinct feedback types. These nine feedback types are the ones we believe to be most relevant for providing and classifying feedback, as they are interpretable, but remain consolidated from human learning research. Finally, one important characteristic of these categories is that they were designed to be as modular as the task allows them to be. This modularity and clear delineation of each feedback class enables the study of feedback types individually, but also of different combinations.

5.3 Mapping Previous NLP Research

To demonstrate the applicability of our taxonomy, we classify feedback examples from prior works according to the nine categories outlined above. If a work employed more than one type of feedback in its examples, we mapped the work to appropriate feedback types. To be best of our knowledge, we are the first to conduct a systematic exploration of textual feedback types to date.

Table 1 presents a cross-referencing of our categorical taxonomy with NLP works employing diverse forms of textual feedback. We observe that the examples of feedback in these works are spread across multiple categories, painting a chaotic picture of how feedback is formulated in current NLP research. While these works broadly consider the *utility* of feedback in how it is *applied* to LLMs, less focus has been given to how the *content* of the feedback affects its utility (and no works ground this content to pedagogical foundations).

Finally, we note that, while we pursue an interpretable categorization to classify these existing works on feedback in NLP, our general taxonomy showcases novel ways to compose feedback, providing an opportunity for further granularity when exploring the space of possible feedback types.

Feedback Type	**Works**
Global Verification	Weston (2016); Shi et al. (2022); Scheurer et al. (2022); Welleck et al. (2022); Wu et al. (2023); Shinn et al. (2023); Chen et al. (2023)
Response Verification	Madaan et al. (2023); Lightman et al. (2023)
Mistakes Verification	Tandon et al. (2022); Saunders et al. (2022); Welleck et al. (2022); Wu et al. (2023); Paul et al. (2023); Chen et al. (2023)
Correct Answer	Weston (2016); Shi et al. (2022); Saunders et al. (2022)
Response Elaboration	Shi et al. (2022); Scheurer et al. (2022); Madaan et al. (2023)
Mistakes Elaboration	Weston (2016); Tandon et al. (2022); Scheurer et al. (2022); Saunders et al. (2022); Shinn et al. (2023)
Task Elaboration	Tandon et al. (2022); Scheurer et al. (2022); Madaan et al. (2023); Chen et al. (2023)
Procedural Elaboration	Weston (2016); Tandon et al. (2022); Shi et al. (2022); Murty et al. (2022); Saunders et al. (2022); Welleck et al. (2022); Schick et al. (2022); Madaan et al. (2023); Shinn et al. (2023)
Metacognition Elaboration	None

Table 1: The distribution of NLP textual feedback research as per the categorical taxonomy, highlighting that categorization remains an unexplored area of feedback research. *n.b.*: Works in grey collect the indicated types of textual feedback, but do not employ it.

6 Case Study

Having established an LLM-relevant feedback model at three levels of abstraction — a feedback ecosystem (§4), feedback content dimensions (§5.1), and concrete feedback categories (§5.2) — in this section we seek to validate its importance by conducting a simple case study.

Using a corpus of media articles written about research papers,[12] we task GPT-4 (OpenAI, 2023)

[12]Details about the dataset, model hyperparameters, and

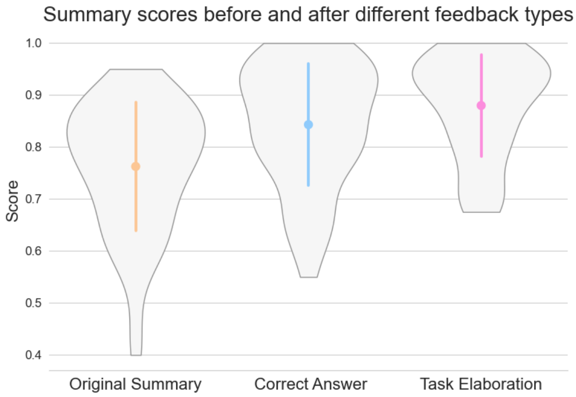

Figure 6: The impact of two different feedback types on a lay summary generation task. Both led to an improvement in average summary quality.

to summarize 50 articles on research papers in lay terms. Then, we provide 2 different types of feedback from our categorization: *Correct Answer* (CA) and *Task Elaboration* (TE), chosen due to their significantly different characteristics. We evaluate both the original and revised answer.[13] More details about the experiment setup and execution are available in Appendix D.

Results Figure 6 demonstrates that, on average, both types of feedback improve the generated summaries. Interesting, we find that *task elaboration* feedback outperforms *correct answer* feedback, which is remarkable, as feedback and evaluation in NLP are usually based on knowledge of the correct answer. However, in this case, feedback that reiterated or provided more information about desired summary qualities (i.e., TE) emerged as the more *useful* of the two feedback types.[14]

Furthermore, the effect of these two feedback types emerges clearly when analyzed qualitatively. For example, TE feedback led to more promotional language — frequently including words such "groundbreaking," "breakthrough," and "revolutionize," which was encouraged by the task elaboration prompt, "A good and captivating summary should first grab the reader's attention" (see Appendix D). Meanwhile, CA feedback often led to more technical language. While this language was likely to be found in some quantity in the *correct answer* (i.e., the gold summary), it was amplified by the model's response, leading to lower scores. Output examples are presented in Appendix D.

Feedback Cost Both CA and TE are answer-independent, (i.e., both are constructed independently from the content of the original answer). Yet, both led, on average, to a significant improvement over the initial summary, demonstrating clear advantages to providing models with feedback. TE, especially, is a simple type of feedback to generate, needing only to describe the task or relevant concepts with some level of detail. The positive impact of such straightforward additions to the generative pipeline reinforces the motivation for studying the space of feedback content more deeply.

7 Conclusion

In this paper, we survey the most influential feedback models from learning science and adjacent social sciences, and create a novel framework and taxonomy of feedback for NLF. Both the framework and taxonomy are constructed to enable a systematic exploration of feedback content, allowing for objective conclusions on the impact of a piece of feedback and the optimization of NLF. The case study (§6) provides early signals into the merits of our formulation. By clearly delineating the information present in a piece of feedback, we can better make inferences about a model's knowledge.

Beyond a survey of the space of natural language feedback, we present a clear path forward, through the deliberate exploration of these feedback categories, the inspection of the impact of varying the more general axes, and the study of the more diverse and peripheral aspects of the FELT framework — such as learner characteristics (different models and different models' parameters), and how they might impact the output.

Future Work Our work opens up several future research directions. One is a more complete validation of this taxonomy, possibly employing carefully LLM-generated feedback to ensure content restrictions are respected. A more interesting question pertains to the study of feedback that RLF methods would consider effective – when using an RLF framework to optimize for feedback, what characteristics emerge? Finally, another open question is the design of feedback for an unknown task, guided by the FELT framework.

prompts used are presented in Appendix D.

[13]As this analysis was conducted on a total of 50 samples, producing 150 evaluations (one per feedback type plus the original summary for each data sample). The evaluation of each summary was conducted by one of the authors, according to an established set of criteria.

[14]This is not surprising from the learning sciences perspective — as seen in Section 3, CA does not possess any of the characteristics of effective feedback.

Limitations

FELT, our proposed framework to capture the full feedback ecosystem, is theorectically grounded and fairly unexplored. Some aspects, like the impact of timing, need to be reassessed for LLMs.

Our previous NLP research mapping (presented in Table 1) was restricted to only the nine feedback content categories. Though they are cleanly delineated and encapsulate the main feedback types that emerged in pedagogical research, they do not encompass all the feedback space captured by the 10 general axes in the taxonomy.

Our case study was conducted exclusively in English. Models trained on other languages, especially lower-resource ones, might react to feedback differently, and it might have a weaker impact on revised generations.

Additionally, we used GPT-4, a paid and closed-source LLM, whose details about architecture and training remain unknown.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: Harmlessness from AI Feedback. ArXiv:2212.08073 [cs].

Robert L. Bangert-Drowns, Chen-Lin C. Kulik, James A. Kulik, and MaryTeresa Morgan. 1991. The Instructional Effect of Feedback in Test-like Events. *Review of Educational Research*, 61(2):213–238. Publisher: [Sage Publications, Inc., American Educational Research Association].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Deborah L. Butler and Philip H. Winne. 1995. Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, 65(3):245–281. Publisher: [Sage Publications, Inc., American Educational Research Association].

David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325. Publisher: Routledge _eprint: https://doi.org/10.1080/02602938.2018.1463354.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Carol Evans. 2013. Making sense of assessment feedback in higher education. *Review of Educational Research*, 83:70–120. Place: US Publisher: Sage Publications.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. ArXiv:2305.00955 [cs].

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. Simulating bandit learning from user feedback for extractive question answering. *arXiv preprint arXiv:2203.10079*.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Avraham N. Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. *arXiv preprint arXiv:2302.08582*.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Raymond W. Kulhavy. 1977. Feedback in Written Instruction. *Review of Educational Research*, 47(2):211–232. Publisher: [Sage Publications, Inc., American Educational Research Association].

Raymond W. Kulhavy and William A. Stock. 1989. Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, 1(4):279–308.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016. Dialogue learning with human-in-the-loop.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue Learning With Human-In-The-Loop. ArXiv:1611.09823 [cs].

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Anastasiya A. Lipnevich, David A. G. Berg, and Jeffrey K. Smith. 2016. Toward a Model of Student Response to Feedback. In *Handbook of Human and Social Conditions in Assessment*. Routledge. Num Pages: 17.

Anastasiya A. Lipnevich and Ernesto Panadero. 2021. A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, 6.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2833–2861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

B. Mason and Roger Bruning. 2001. Providing Feedback in Computer-based Instruction: What the Research Tells Us. *Center for Instructional Innovation*, 15.

Santosh Mathan and K. Koedinger. 2005. Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, 40:257 – 265.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese.

2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Shikhar Murty, Christopher D. Manning, Scott Lundberg, and Marco Tulio Ribeiro. 2022. Fixing model bugs with natural language patches. *arXiv preprint arXiv:2211.03318*.

Susanne Narciss. 2008. Feedback Strategies for Interactive Learning Tasks. pages 125–144.

Susanne Narciss and Katja Huth. 2004. How to design informative tutoring feedback for multi-media learning.

David J. Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218. Publisher: Routledge _eprint: https://doi.org/10.1080/03075070600572090.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ernesto Panadero and Anastasiya A. Lipnevich. 2022. A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35:100416.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Arkalgud Ramaprasad. 1983. On the Definition of Feedback. *Behavioral Science*, 28:4–13.

D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback. *arXiv preprint arXiv:2204.14146*.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*.

Weiyan Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. ArXiv:2009.01325 [cs] version: 3.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 339–352, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Jason Weston. 2016. Dialog-based Language Learning. ArXiv:1604.06045 [cs].

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2211.00053*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *arXiv preprint arXiv:2208.03270*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Pedagogical models of feedback

A.1 Defining feedback

Table 2 presents an overview of the various definitions of feedback put forward by several pedagogical works.

A.2 Categorizing feedback

Kulhavy and Stock (1989) model feedback as having two components: the verification component, f_v, which is a simple discrete classification of the answer as correct or incorrect, and the elaboration component, f_e, consists of three elements:

(i) *type*, whether the feedback contains information derived from the current task (task-specific), not from the task but from the relevant lesson (instruction-based), or beyond the relevant lesson, such as new information, examples or analogies not previously introduced (extra-instructional),

(ii) *form*, the difference in structure between the feedback and instruction or task specification messages, requiring increased processing the less similar it is[15], and

(iii) *load*, the total amount of information in the feedback - from a single "correct/incorrect" bit to including the correct answer to even more informative feedback accompanying it with an explanation, for example.

Mason and Bruning (2001) propose 8 feedback categories, arguing different types of feedback are best suited for different learner characteristics, taking into account the students' proficiency and prior knowledge, as well as the task difficulty. The eight categories are:

(i) *no-feedback*, which presents a single grade,

(ii) *knowledge-of-response*, which analogously to the aforementioned verification component, indicates whether the given answer is correct or incorrect,

(iii) *answer-until-correct*, an iterative variant of knowledge-of-response feedback, not allowing the student to progress until they have provided the correct answer,

(iv) *knowledge-of-correct-response*, which provides the correct answer,

(v) *topic-contingent*, which provides both knowledge-of-response feedback and, analogously to Kulhavy and Stock (1989)'s instruction-based type of feedback, provides general information about the topic of the task, where the learner might locate the correct answer,

(vi) *response-contingent*, which similarly provides knowledge-of-response feedback as well as an explanation of why the answer is wrong or right (mapping it to Kulhavy and Stock (1989)'s extra-instructional type of feedback),

(vii) *bug-related*, providing knowledge-of-response feedback and bug-related feedback, which relies on rule sets to identify procedural errors, and

(viii) *attribute-isolation*, which provides knowledge-of-response feedback as well as information on the essential attributes of the relevant concept, focusing the learner on its key components.

Narciss and Huth (2004); Narciss (2008) present a detailed and comprehensive feedback model, taking into account many learner and task characteristics. They also present a content-related feedback classification scheme, with eight categories:

(i) *Knowledge of performance (KP)*, analogous to Mason and Bruning (2001)'s no-feedback and Kulhavy and Stock (1989)'s verification component for a multiple-question task, presents the learner with an aggregate score (e.g., percentage or number of correct answers out of the total number of questions),

(ii) *Knowledge of result/response (KR)*, directly mirrors Mason and Bruning (2001)'s knowledge-of-response and Kulhavy and Stock (1989)'s verification component for each question or task, classifying it as either correct or incorrect,

(iii) *Knowledge of the correct results (KCR)*, equivalent to Mason and Bruning (2001)'s knowledge-of-correct-response, indicating the correct answer to the learner,

(iv) *Knowledge about task constraints (KTC)*, somewhat similar to Mason and Bruning (2001)'s topic-contingent feedback, is elaboration feedback about the task, containing hints, examples or explanations about the type

[15] The *form* element does not apply to *extra-instructional type* feedback, as there is no structural comparison point possible

of task, its rules, sub-tasks, requirements and other constraints,

(v) *Knowledge about concepts (KC)*, containing some resemblance to Mason and Bruning (2001)'s attribute-isolation feedback, is elaboration feedback on the relevant concepts, providing hints, examples or explanations on technical terms, the concept or its context, attributes, or key components,

(vi) *Knowledge about mistakes (KM)*, which parallels Mason and Bruning (2001)'s bug-related feedback, provides elaboration feedback containing the number of mistakes, their location, and hints, examples or explanations on error types and sources,

(vii) *Knowledge about how to proceed (KH)*, elaboration feedback on the general know-how of the task, containing hints, examples or explanations on error correction, task-specific solving strategies or processing steps, guiding questions and worked-out examples, and

(viii) *Knowledge about metacognition (KMC)*, elaboration feedback going beyond the context of the current task, and providing hints, examples, explanations, or guiding questions on metacognitive strategies.

Hattie and Timperley (2007) present a small typology about the information being conveyed about the learner in the feedback message, presenting 3 questions feedback can answer:

(i) where the learner is going (*feed up*),

(ii) how they are going (*feed back*), and

(iii) where to next (*feed forward*)

and argue feedback is effective if it answers all three.

Work	Feedback Definition
Ramaprasad (1983)	Information which changes the gap between "the actual level and the reference level of a system parameter." This is quite a strict definition – if the information does not change the gap, it is not considered feedback, and information about the actual level, the reference level and their comparison is needed beforehand.
Kulhavy and Stock (1989)	Refer to a previous definition of feedback, whereby feedback is considered "any of the numerous procedures that are used to tell a learner if an instructional response is right or wrong" (Kulhavy, 1977).
Sadler (1989)	"Information about how successfully something has been or is being done."
Butler and Winne (1995)	A way to update the learner's internal state and knowledge, and subsequently task execution (a more learner-centric model of feedback).
Kluger and DeNisi (1996)	The information provided by an external agent on one or more aspects of task performance. Note this excludes the learner as a possible source of feedback.
Mason and Bruning (2001)	Feedback "is any message generated in response to a learner's action."
Narciss and Huth (2004); Narciss (2008)	"All post-response information which informs the learner on his/her actual state of learning or performance in order to regulate the further process of learning in the direction of the learning standards strived for."
Nicol and Macfarlane-Dick (2006)	Information relating the learner's current state to the goal state (both with regards to learning as well as performance). Importantly, they consider students generate internal feedback and that the better they are at self-regulation, the better they will be at either generating or leveraging feedback.
Hattie and Timperley (2007)	Information generated by an agent about the learner's understanding or their performance.
Evans (2013)	Feedback "includes all feedback exchanges generated within assessment design, occurring within and beyond the immediate learning context, being overt or covert (actively and/or passively sought and/or received), and importantly, drawing from a range of sources."
Lipnevich et al. (2016)	Feedback is information transmitted to the learner with the intent of changing their understanding and execution, in order to improve learning.
Carless and Boud (2018)	Feedback as the process through which the student understands and integrates information from various sources in order to improve their learning or performance (a more learner-centric perspective).
Lipnevich and Panadero (2021)	Feedback "is information that includes all or several components: students' current state, information about where they are, where they are headed and how to get there, and can be presented by different agents (i.e., peer, teacher, self, task itself, computer). This information is expected to have a stronger effect on performance and learning if it encourages students to engage in active processing."

Table 2: Different pedagogical works' definitions of feedback.

B The FELT framework's components

The FELT framework introduced in Section 4 presents an important overview of all the factors that influence feedback and are in turn influenced by it. Figure 4 showcased a schematic overview of the FELT framework, integrating four distinct components: Feedback, Errors, Learner, and Task. In this appendix, we will outline more precisely each of the components of the FELT Framework, as well as the interactions between them.

B.1 Task

Typically, the task will be the first element to be defined.

Nature of Task In this paper, we have limited tha nature of the task to the answer type. Understanding itis fairly easy – a task has a closed-answer if there is a finite set of correct answers, and an open-answer otherwise. Notably, tasks can contain both elements. For example the task "*Write a quality 4-paragraph short-story*" has both open- and closed-answer elements. There is no finite set of answer of what a quality story is, but whether a story has 4 paragraphs, or not, is a binary closed-answer, as seen in Appendix C.

Complexity The difficulty level of a task is harder to define as some measure of relativity is involved. We suggest anchoring this measurement to the average adult human capabilities. A simple arithmetic task will thus be considered very easy, whereas researching and writing a doctoral thesis would be seen as hard.

Prompt Instructions The task instructions will be presented to the model at two distinct points in time: when first assigning the model this task, and when later providing feedback. With regards to the former, this element captures the degree to which the task is explained – is the model explicitly aware of all criteria it should satisfy? With regards to the second pass, when feedback is provided, this dimension pertains instead with the degree of freedom it gives the LLM – is the model forced to take the feedback into account, or can it consider only part of it, or even disregard it altogether if it deems it useless?

B.2 Learner

Either at the same time the task is defined or immediately after, the model to be tested will be chosen. The model choice influences two important features.

Prior Knowledge The prior knowledge captures the LLM's abilities as a direct result of its size, training data, and training method. These, in turn, also reflect the model's purpose (e.g., was it designed to be helpful, harmless, entertaining, etc.). The prior knowledge thus captures the model's representation of the learner, and in its architecture and parametric knowledge, it encodes the LLM's current abilities – or its proficiency – both in general and with regards to the specific task.

Feedback Processing Mechanism Mainly defined by the experimental setup, the mechanism by which the model process feedback can vary significantly, and not all of them are able to leverage the same level of information. Imitation learning, for example, can only leverage information which was positively evaluated. As stated in Section 4, we identify 4 main processing mechanisms, 3 of which alter the model's parametric state – feedback-based imitation learning, joint-feedback modeling, and reinforcement learning, as defined in Fernandes et al. (2023) – and a fourth, non-parametric mode: in-context learning (Brown et al., 2020).

B.3 Errors

After both the task and learner are in place, the first pass of the experiment can be run, where the model will have its first attempt at solving the task. In this attempt, it is expected that the model will make some degree of mistakes – which have two important characteristics.

Error Type There are several possible types of errors, and their differences are significant. For example, an error made due to a guess only needs to provide the learner with the right information for it be be corrected, whereas a systematic error (for example, the mixing of British and American English spellings) will require a different, much more insistent, intervention. ROSCOE (Golovneva et al., 2023) proposes a taxonomy of step-by-step reasoning errors. While task dependent (i.e., there are grammar errors and arithmetic errors, rather than fully task independent failure modes), this taxonomy provides a good starting ground for the exploration of error types in NLP.

Error Severity Besides the type of error, it is also important to take the severity of the error into account. Stating that Marie Curie was a German philosopher and stating that she won one Nobel Prize in her lifetime are both factually inaccurate – but one is a severe, complete hallucination, while the other omitted she actually won the Nobel Prize twice. The more severe the error, the stronger, more insistent, and more corrective the feedback should be.

B.4 Feedback

Finally, after the model has finished its first attempt at the task, producing some number of errors, feedback can be provided on this attempt.

Timing One easy to neglect aspect of feedback that pedagogy has shown to be impactful is timing – whether the feedback is provided immediately after a task is attempted or whether there is a delay between the two actions. There are differing opinions amongst education researchers, but how to make feedback content more effective through timing merit research in LLMs. For example, in line with Mathan and Koedinger (2005) and Narciss (2008)'s take on timing – delay feedback if the learner possesses metacognitive abilities that allow them to identify and possibly correct mistakes – we posit feedback will be more effective if, content-wise, it is preceded by information on the answer's correctness and mistakes' existence and only after this metacognitive priming is the rest of the information presented.

Content Section 5 explores feedback content in depth, presenting 10 impactful axes on which it can vary: length, granularity, applicability of instructions, answer coverage, criteria, information novelty, purpose, style, valence, and mode. It also presents a set of 9 emergent categories which, based on pedagogical research, we estimate to be the most promising one with regards to impact on revised model generations, and thus most deserving of further study.

Source Finally, it is also important to consider the source of feedback, which might be an authority, such as an expert, an average human, another LLM, a rule-based system, among others. Different sources will reflect different authority and reliability levels.

B.5 Interactions

With a clear understanding of all the components and sub-components of the FELT framework, we can explore the influences that exist between them.

Both the task complexity and the learner's prior knowledge can impact the ideal feedback timing – be it delayed when the learner has metacognitive skills (Narciss, 2008) or enough task proficiency (Mason and Bruning, 2001) they can identify where the mistake occurred, or, for example, immediate if they don't (Narciss, 2008) or the task difficulty is low (Mason and Bruning, 2001).

With regards to the feedback content, the type of task (Butler and Winne, 1995; Kluger and DeNisi, 1996; Mason and Bruning, 2001; Lipnevich et al., 2016) and both the error type and severity will have an impact (Narciss and Huth, 2004; Narciss, 2008). The nature of the task (open or closed answer) will directly condition the feedback that can be given in response to the model's answer, as well as how difficult it will be to produce it. For example, generating the correct answer for a multiple choice quiz or a story writing task will be two very different endeavors. Similarly, it is impossible to provide response elaboration feedback on a single multiple choice question. The error type and severity will also influence the feedback content, as apart from directly dictating what mistakes verification and elaboration feedback can be given, they will also condition the ideal amount of detail and explanations to address the mistake at the most efficient level.

Finally, all aspects of feedback will influence the learner's feedback processing mechanism (Kulhavy and Stock, 1989; Sadler, 1989; Bangert-Drowns et al., 1991; Butler and Winne, 1995; Kluger and DeNisi, 1996; Narciss and Huth, 2004; Nicol and Macfarlane-Dick, 2006; Narciss, 2008; Lipnevich et al., 2016; Carless and Boud, 2018). All three dimensions of feedback have evident potential to directly influence how the model processes them. The instruction's permissiveness to consider or discard feedback will also impact the learner's feedback processing mechanism. This processing is, of course, dependent on the specific processing mechanism employed, and while some might be indifferent to some of these components – like imitation learning, for example, which focuses exclusively on the feedback content – others will be sensitive to all, including the task's prompt instructions – such as in-context learning.

C Different feedback content types

In this appendix, we first present different examples of the nine feedback content categories for the summarization task used in the case study. Second, we present a more detailed description of each feedback type, to facilitate feedback categorization according to the categorical taxonomy.

C.1 Task Introduction

In this task, the model is instructed to summarize a news article about a research finding into a summary that can be understood by an adult who is not particularly familiar with the field. As more thoroughly described in Appendix D, the summary should be engaging while also describing the problem, the finding and the venue of publication.

C.2 Feedback Examples

Let us consider the following generated summary:

`"Scientists at the University of Bern have used 3D computer simulations to reconstruct how protoplanet Vesta collided with other asteroids around a billion years ago, using data from NASA's 2011 Dawn probe. The simulations also reveal detailed information about Vesta's composition and properties, adding to our understanding of the solar system's evolution. The findings were presented in a study published in Nature magazine."`

Below is a possible example of each type of feedback that could be given in response:

1. **Global Verification:** Overall grade of 85%.
2. **Response Verification:** Describes the approach and the motivation, and mentions the venue, but not particularly engaging and incomplete about authorship.
3. **Mistakes Verification:** No initial attempt at grabbing the reader's attention; inaccurate authorship attribution (incomplete).
4. **Correct Answer:** The correct answer is: Models boost the significance of image and measurement data from space missions and help to understand our solar system. A simulation of a double impact that occurred on the proto-planet Vesta one billion years ago allowed scientists to describe precisely the inner structure of the asteroid. A joint research from EPFL, Bern University, France and the United States is on the cover of Nature this week.
5. **Response Elaboration:** The summary is accurate overall, and the language employed is adequate for the target audience. It could, however, use improvement in several areas, such as being more attention grabbing, especially in its first sentence, and providing more detail about what was actually found, rather than generic phrases like "adding to our understanding of the solar system's evolution.". It also failed to capture it was a joint research project.
6. **Mistakes Elaboration:** The summary does not attempt to engage with the user and capture their attention, in order to make them curious to read it to its completion. Furthermore, it only mentions one of the universities collaborating in this research, in which several labs from different universities joined efforts.
7. **Task Elaboration:** A good and captivating summary should first grab the reader's attention and make them curious to learn more. It is then important to factually and precisely state what the problem is, why it is important, the proposed solution, and, if published or being divulged, disclose where the reader can find it.
8. **Procedural Elaboration:** While reading the article, it is important to identify the key aspects of the research – what problem or research question was being studied, what approach was used to solve it, and what its contributions or applications are. It is also important to register who was behind the findings and where they were published. Finally, it is important to think about how to present this idea in the first sentence – in a way that is engaging for the reader, getting their attention and making them curious to read the rest.
9. **Metacognition Elaboration:** To achieve a given task, it is first necessary to understand it and the concepts involved. With this understanding, it is possible to identify the task's goal. Then, one must determine the steps needed to achieve this goal. If needed, each of these steps can be further split into even smaller tasks. Finally, it can be helpful to

establish timelines and deadlines for each of these tasks, so the goal is achieved in time.

C.3 Observations

Information Overlap It is possible to observe some of these feedback types share some information, that is, that there is some overlap between different categories. This is a natural consequence from the task type. Indeed, this summarization task is not a fully open answer question. While there is significant answer space, there are some hard requirements, such as mentioning the research finding, its motivation, the publication venue, being engaging, and using an accessible language. Almost all these are binary requirements, which a summary either fulfills or fails to address. As such, there will naturally be some overlap with regards to them in the feedback information, as it is only possible to fully prevent it in a truly open-answer setting.

Feedback Effectiveness In Section 3, Hattie and Timperley (2007)'s definition of effective feedback was presented. According to the authors, it should address three different information needs: where the learner is going (*feed up*), how they are going (*feed back*), and where to next (*feed forward*). It is possible to relate these questions to the feedback categories exemplified above.

The *feed up* question, that is, the goal performance, can be implicitly derived from all feedback types that describe flaws of the current answer (by exclusion) or its merits (by inclusion). However, it is the *Correct Answer* feedback category that directly and explicitly presents this information.

The *feed back* question – the learner's current performance – is derived from all the verification feedback categories as well as from *Response Elaboration* and *Mistakes Elaboration*.

Finally, the *feed forward* question, how the learner should proceed, is directly tackled by *Procedural Elaboration* feedback.

This definition would then, disregard *Task Elaboration* and *Metacognition Elaboration* feedback categories as inefficient feedback. However, as the case study presented in Section 6 demonstrates, at least in the NLP domain, we cannot be so quick to dismiss them – as in it, *Task Elaboration* actually outperformed *Correct Answer* feedback. Consequently, while the definition of effective feedback proposed by Hattie and Timperley (2007) might help researchers consider promising feedback types, it might also dismiss other pertinent pieces of information. For that reason, both *Task Elaboration* and *Metacognition Elaboration* are present in the categorical taxonomy proposed by this paper, despite their lack of clear mapping to any of Hattie and Timperley (2007)'s three questions.

C.4 Mapping Feedback to a Categorical Type

While examples of the nine feedback categories might be easy to understand, classifying novel pieces of feedback might prove challenging. Below, we provide a more exhaustive overview of the content of each type of feedback:

1. **Global Verification:** An aggregate score for the task as a whole. Cannot contain more than a single data point of information. The score need not be numeric (e.g., "`Satisfactory`," "`Grade: 65%`," and "`C`" are all valid examples of Global Verification feedback).

2. **Response Verification:** Granular response-classification feedback. Can either provide a score for several answer segments (e.g., a unique score for each question on a quiz, or for each paragraph on a written text) or a score for several evaluative criteria (e.g., evaluate the entire written text on readability, engagement, etc.).

3. **Mistakes Verification:** Granular error-classification feedback. Can either simply state errors were committed or identify which types of errors are present in the submitted answer (it can mention the number of mistakes).

4. **Correct Answer:** The correct answer, an expected solution or, for an open-ended task, a rewritten version of the submitted answer that fulfills the evaluative criteria (ideally with as few changes as possible).

5. **Response Elaboration:** An overview of the answer as a whole, incorporating feedback about the current level of performance of the student. It can choose to only mention part of it, focusing only on the learner's positive accomplishments or their shortcomings. It differs from mistakes elaboration as, though it can discuss shortcomings, it does not directly address mistakes.

6. **Mistakes Elaboration:** Detailed feedback about mistakes, including their location, thorough descriptions of their type and of their possible sources (e.g., information about common mistakes and what misconceptions might lead to them). Can either be explicit or presented through hints or guiding questions. Note that no information on correcting mistakes is included as part of this feedback type, as these belong to the Procedural Elaboration feedback instead.

7. **Task Elaboration:** Clarifications about the task — its type, requirements, constraints, subprocesses — and relevant concepts and technical terms. Can either be explicit or presented through hints or guiding questions. Note, however, no information on task solving strategies is considered Task Elaboration type of feedback, as these belong to the Procedural Elaboration feedback instead.

8. **Procedural Elaboration:** Instructions on how to improve performance, be it through worked out examples, explanations on error correcting, or strategies for processing and solving the task. Can either be explicit or presented through hints or guiding questions.

9. **Metacognition Elaboration:** General strategies for learning and problem solving. This feedback cannot be directly related to the task being attempted by the learner. Can either be explicit or presented through hints or guiding questions.

D Case Study Implementation Details

D.1 Dataset and Model

The task presented in section 6 involved the summarizing of a news article describing a research development at EPFL in more approachable language and terminology. The summary had as an objective to be even more approachable to people outside the field.

Dataset The data for this task came from EPFL's Mediacom department, where they provided the authors with a set of 2370 entries of articles, summaries, and extra information (title, author, date, etc.). Out of these, 50 were chosen so that all articles relayed a newly published work. This was the only criteria for selection. All articles and summaries were pre-processed so as to remove HTML tags.

Model The model used for this task was GPT-4 (OpenAI, 2023), with its default hyperparameters, called through OpenAI's Chat Completions API. Thus, the model generated a single completion for each prompt, with a temperature of 1, and with no limitation on the maximum number of tokens (beyond the model's own context length).

D.2 Experiment Execution

The experiment was run in two stages. In the initial phase, the model is asked to generate a first summary. It is then provided with feedback and asked to revise its original summary. In this experiment, two distinct types of feedback were provided: *Task Elaboration* (TE) and *Correct Answer* (CA).

Initial Generation Following OpenAI's Chat Completions API, the model prompting is done under a chat format. In this setting, the first *message* is a system message stating `You are a helpful assistant`. This is then followed by an user message, with the following prompt:

```
Summarize the following article into a short but captivating snippet under around 100 tokens. It must describe both the problem and the approach used to solve it, as well as the venue where these findings were presented, whenever this information is available.
Article: [article body]
```

The model's response message to this prompt is considered its original summary.

Revised Generation The revised generation prompt is shares the chat prompting format. It contains the previous chat history, which includes not only the two messages outlined above but also the model's answer as an assistant message. To these three messages, a new user message is added, with the following content:

```
Feedback: [feedback]
Please revise your original summary taking the feedback into consideration. If you feel the feedback is not appropriate or useful, you can disregard it.
```

The `[feedback]` placeholder will have one of two different values, depending on the feedback type being provided:

- **Correct Answer:** The feedback will be of the form
  `The correct answer is: [gold_summary]`
  where `[gold_summary]` is the summary provided by the Mediacom dataset,

- **Task Elaboration:** The feedback will be of the form
  `A good and captivating summary should first grab the reader's attention and make them curious to learn more. It is then important to factually and precisely state what the problem is, why it is important, the proposed solution, and, if published or being divulged, disclose where the reader can find it.`

Finally, as in the first stage, the model's response message to this prompt is considered its revised summary.

D.3 Example Outputs

In this subsection, we present a few examples outputs from the case study.

D.3.1 Example 1

Original Article `The International Consortium of Investigative Journalists (ICIJ), which has over 200 members in 70 countries, has broken a number of important stories, particularly ones that expose medical fraud and tax evasion. One of its most famous investigations was the Panama Papers, a trove of millions of documents that` revealed the existence of several hundred thousand shell companies whose owners included cultural figures, politicians, businesspeople and sports personalities. To complete an investigation of this size is only possible through international cooperation between journalists. When sharing such sensitive files, however, a leak can jeopardize not only the story's publication, but also the safety of the journalists and sources involved. At the ICIJ's behest, EPFL's Security and Privacy Engineering (SPRING) Lab recently developed Datashare Network, a fully anonymous, decentralized system for searching and exchanging information. A paper about it will be presented during the Usenix Security Symposium, a worldwide reference for specialists, which will be held online from 12 to 14 August.

Anonymity at every stage

Anonymity is the backbone of the system. Users can search and exchange information without revealing their identity, or the content of their queries, either to colleagues or to the ICIJ. The Consortium ensures that the system is running properly but remains unaware of any information exchange. It issues virtual secure tokens that journalists can attach to their messages and documents to prove to others that they are Consortium members. A centralized file management system would be too conspicuous a target for hackers; since the ICIJ does not have servers in various jurisdictions, documents are typically stored on its members' servers or computers. Users provide only the elements that enable others to link to their investigation.

Users searching for information enter keywords in the search engine. If the search produces hits, they can then contact colleagues – whose identity remains protected – who are in possession of potentially relevant documents. Search queries are sent encrypted to all users, if there is a macth the querier gets an alert and can decide whether they wish to enter in contact and share information. "Given the fact that users work in different time zones, some with only a few hours of internet access per day, it was critical that searches and responses could take place asynchronously," notes Carmela Troncoso, who runs the SPRING Lab at the School of Computer and Communication Sciences (IC). Another messaging system, also secure and anonymous, is subsequently used for two-way exchanges.

Two completely new secure applications

"This system, which addresses real-world needs, has enabled SPRING to tackle some interesting challenges," notes Troncoso. The research team drew on existing authentication mechanisms and anonymous communication primitives, which they then optimized. They also developed two completely new secure building blocks that are described in the paper being published today in connection with the Usenix Security Symposium – an asynchronous search engine and a messaging system. A new protocol, known as "multi-set private set intersection" (MS-PSI), ensures the security of the search engine, allowing users to easily search a large number of databases without increasing the risk of leaks. The messaging system relies on a large number of single-use virtual mailboxes and is based on the well-known "pigeonhole" system, which chooses one option at random, in this case one of the mailboxes. Currently, the system does not allow users to exchange documents. "At this stage in the process, journalists are using other secure messaging systems," Troncoso says.

Working with the Consortium has allowed SPRING to frame new requirements that are rarely examined in the scientific literature. Datashare can be scaled to thousands of users and millions of documents while encrypting all communications. "The hurdles we encountered during the development process, however, have paved the way to a new area of research with significant potential for other fields," Troncoso concludes.

Golden Summary An EPFL laboratory has developed Datashare Network, a decentralized search engine paired with a secure messaging system that allows investigative journalists to exchange information securely and anonymously. An scientific article on this subject will be presented during the Usenix Security Symposium which will be held online from August 12 to 14.

GPT4 Original Summary EPFL's Security and Privacy Engineering Lab developed an anonymous, decentralized system called Datashare Network to aid the International Consortium of Investigative Journalists securely exchange information. Offering user anonymity, this system will be presented at the Usenix Security Symposium held online from 12-14 August.

***Correct Answer* Revision** EPFL's Security and Privacy Engineering Lab created Datashare Network, a decentralized search engine with a secure messaging system that enables investigative journalists to exchange information anonymously and securely. A scientific paper on this solution will be showcased at the online Usenix Security Symposium from August 12 to 14.

***Task Elaboration* Revision** Investigative journalism goes undercover with Datashare Network, a breakthrough anonymous, decentralized information-sharing system. Developed by EPFL's Security and Privacy Engineering Lab, the network protects journalists from exposure while tackling global issues like tax evasion and medical fraud. Discover how it secures data-sharing at the Usenix Security Symposium, held online from 12-14 August.

D.3.2 Example 2

Original Article Thermoelectric materials in particular hold vast potential for use in energy applications because they generate electricity from waste heat, such as that generated by industrial processes, by car and truck engines, or simply by the sun. Reducing the thermal conductivity of these materials by a factor of three, for example, would completely revolutionize existing waste-heat recovery, and also all refrigeration and air-cooling technology.
A unique theory for all insulating materials
In the paper Unified theory of thermal transport in crystals and glasses, out in Nature Physics, Michele Simoncelli, a PhD student at EPFL's Theory and Simulation of Materials (THEOS) Laboratory – together with Nicola Marzari, a professor at EPFL's School of Engineering and head of THEOS and of the MARVEL NCCR, and Francesco Mauri, a professor at the University of Rome–Sapienza – present a novel theory that finally decodes the fundamental, atomistic origin of heat conduction. Up to now, different formulations needed to be used depending on the systems studied (e.g., ordered materials, like a silicon chip, or disordered, like in a glass), and there wasn't a unified picture covering all possible cases.
This has now been made possible by deriving directly from the quantum mechanics of dissipative systems a transport equation that covers on equal footing diffusion, hopping, and tunneling of heat. Waste heat recovery
This fundamental understanding will allow scientists and engineers to accurately predict the thermal conductivity of any insulating material (in metals, the heat is carried by the electrons, and that is well understood) – this is exceedingly important for thermoelectrics (i.e. materials that can convert heat into electricity), since these have both crystal- and glass-like properties, and are much needed for waste-heat recovery, or for refrigeration without greenhouse gases (and if you think refrigeration is boring, it is worth remembering that Albert Einstein spent many years trying

to invent a new form of refrigerator).
In order to develop such next-generation technology, however, scientists first need to understand how and to what extent materials conduct heat. "Up to now, two different equations have been used for calculating thermal properties: one describes perfectly crystalline materials – that is, materials with highly ordered atomic structures – and the other one completely amorphous materials like glass, whose atoms do not follow an ordered pattern," says Michele Simoncelli. These equations happened to work well in those special cases. "But between these two extremes lie a plethora of interesting cases, and neither equation worked – this is really where our contribution makes a profound difference"
Illustration : Heat is the energy due to vibrations of atoms. The ridges represent the quantum vibrational energy levels of CsPbBr3, a promising thermoelectric material. Heat conduction originates from both diffusion of vibrational excitations (blurred spheres, following the energy level) and quantum tunnelling (jump between different energy levels, represented by blue waves).

Golden Summary EPFL and MARVEL researchers have developed a new theory for heat conduction that can finally describe and predict the thermal conductivity of any insulating material. This new formulation will let scientists make accurate predictions of thermal conductivity in a range of materials for critical engineering applications – from electronics to lasers to waste-heat recovery.

GPT4 Original Summary Scientists at EPFL's Theory and Simulation of Materials Laboratory have developed a unified theory to predict the thermal conductivity of insulating materials, opening up possibilities for more efficient energy applications. The research breakthrough, published in Nature Physics, could lead to waste-heat recovery, revolutionising refrigeration and air-cooling technology without greenhouse gases. Understanding how materials conduct heat is key to developing next-generation thermoelectric technology.

***Correct Answer* Revision** Researchers at EPFL and MARVEL have developed a groundbreaking theory for heat conduction, enabling the description and prediction of thermal conductivity in any insulating material. Published in Nature Physics, the new formulation offers potential for critical engineering applications such as waste-heat recovery, electronics, and lasers, enhancing energy efficiency in various industries.

***Task Elaboration* Revision** Researchers have unlocked the secret to more efficient energy applications by developing a unified theory for predicting the thermal conductivity of insulating materials, transforming waste-heat recovery and revolutionizing refrigeration technology. With this innovative approach, published in Nature Physics, scientists can now accurately predict and optimize thermoelectric materials, paving the way for energy-efficient solutions without greenhouse gas emissions. Discover the full potential of this game-changing breakthrough at EPFL's Theory and Simulation of Materials Laboratory.

D.3.3 Example 3

Original Article With its 500 km diameter, the asteroid Vesta is one of the largest known planet embryos. It came into existence at the same time as the Solar System. Spurring scientific interest, NASA sent the Dawn spacecraft on Vesta's orbit for one year between July 2011 and July 2012.
Data gathered by Dawn were analyzed by a team of researchers from EPFL as well as the Universities of Bern (Switzerlanf), Brittany (France) and Arizona (USA). Conclusion : the asteroid's crust is almost three times thicker than expected. The study does not only have implications for the structure of this celestial

object, located between Mars and Jupiter. Their results challenge a fundamental component in planet formation models, namely the composition of the original cloud of matter that aggregated together, heated, melted and then crystallized to form planets.

At EPFL's Earth and Planetary Science Laboratory (EPSL), led by Philippe Gillet, Harold Clenet had a look at the composition of the rocks scattered across Vesta's ground. "What is striking is the absence of a particular mineral, olivine, on the asteroid's surface," said the researcher. Olivine is a main component of planetary mantles and should have been found in large quantities on the surface of Vesta, due to a double meteorite impact which, according to computer simulations, "dug" the celestial body's southern pole to a depth of 80 km, catapulting large amounts of materials to the surface.

The two impacts were so powerful that more than 5 % of Earth's meteorites come from Vesta. « But these cataclysms were not strong enough to pierce through the crust and reach the asteroid's mantle,» Harold Clenet continued. The meteorites originating from Vesta and found on Earth confirm this since they generally lack Olivine, or contain only minute amounts compared to the amount observed in planetary mantles. Also, the spacecraft Dawn did not find olivine in the vicinity of the two impact craters. « This means that the crust of the asteroid is not 30 km thick, as suggested by the models, but more than 80 km. »

Composition of planets

These discoveries challenge models that describe the formation of Vesta, and consequently the formation of rocky planets in the Solar System including planet Earth. Cooling theory and "re-melting" phenomena in the depths of previously solidified elements would also need to be reviewed. "The crust might have been thickened by the formation of "plutons", that is: igneous rock intrusions, hundreds of meters large, some of which emerged to the surface," explained the scientist.

If Vesta has less of a (olivine-rich) mantle and more of a (pyroxene-rich) crust, then the proportion of materials making up Vesta, and probably the Earth and other telluric planets (Mars, Venus, Mercury), is different from what was previously expected.

A more complex model of planet formation therefore has to be considered, one that takes into account not only the original composition of planets, but also their orbits, sizes and related cooling times. Vesta is the only known asteroid that has an earth-like structure - with a core, mantle and crust – making it an incredible laboratory for testing hypotheses and theories.

Golden Summary EPFL researchers have a better understanding of the asteroid Vesta and its internal structure, thanks to numerical simulations and data from the space mission Dawn. Their findings, published today in Nature, question contemporary models of rocky planet formation, including that of Earth.

GPT4 Original Summary Data collected from asteroid Vesta poses challenges to planet formation models as it reveals the asteroid's crust is almost three times thicker than expected. The discovery, made by researchers from the Earth and Planetary Science Laboratory and partner universities, suggests that original assumptions about the composition of the cloud of matter that formed planets may be incorrect. The findings could lead to a more complex model of planet formation, taking into account composition, orbit, size, and cooling times.

***Correct Answer* Revision** EPFL researchers have gained insights into asteroid Vesta's internal structure, utilizing numerical simulations and data from the Dawn space mission. The findings, published in Nature, indicate Vesta's crust is almost three times thicker than expected, questioning current models of rocky planet formation, including Earth's. This could lead to more complex planet formation models that consider composition, orbit, size, and cooling times.

***Task Elaboration* Revision** Astonishing findings from asteroid Vesta are challenging our understanding of planet formation, as researchers discovered its crust is nearly triple the expected thickness. This breakthrough, made by EPFL's Earth and Planetary Science Laboratory and partner universities, questions key elements of planet formation models and paves the way for a more complex approach, considering composition, orbits, sizes, and cooling times. With Vesta being the only known asteroid with an Earth-like structure, this revelation creates a compelling lure for further explorations.