Reconstruction template:

1. If all multi-same arguments are in the same subargument, we can simply use connection words like **“and”.**

2. If not all multi-same arguments are in the same subargument, then it implies some subarguments are supplement to one subarguments, e.g., evidence/support supplement to observation in **Result**; implementation supplement to methodology in **Method**.
   In this scenario, we need to identify the subargument for each **while doing reconstruction**. Then use one indicator words for the subargument and link them fluently together. E.g., the Result show (observation), supported by (evidence); the Method was conducted using (methodology) which is developed with (implementation)…

Below is a complete example template that include all subarguments:

Agent + Main Action + Object + building on (**Historical Context**) while assuming that (**Assumption**) building on (**Historical Context**) while assuming that (**Assumption**), under specific conditions such as (**Condition**), and grounded in (**Theoretical Framework**). It differentiates itself by (**Differentiation**). The goal of this work is to (**Goal/Aim**), tested by hypothesizing that (**Hypothesis**). The methodology employed includes (**Methodology**), which involves steps like (**Algorithm/Process Description**), evaluated against (**Benchmarking**). The implementation is developed using (**Implementation**), refined through (**Optimization**), and functions as described in (**Mechanism**). The results demonstrate (**Observation**), supported by (**Evidence**), and categorized as (**Classification/Categorization**). This work introduces (**Novelty**) and outperforms others in (**Comparison**). The findings are validated through (**Validation**), measured with (**Metrics**), and shown to scale effectively via (**Scalability**). The results highlight relationships such as (**Causation**) or (**Correlation**), while challenges like (**Limitation**) and errors such as (**Error**/**Artifact**) are acknowledged within the findings. Observed outcomes or actions are explained by (**Explanation**). Ethical concerns include (**Ethical Risks**), but these are addressed through (**Ethical Justification**), benefiting (**Stakeholders/Beneficiaries**). The implications suggest (**Implication**), despite risks such as (**Risk/Uncertainty**), with future directions addressing (**Future Work/Questions**). Finally, this work challenges prior knowledge by presenting (**Contradiction**), considers (**Alternative/Counterargument**), and defends against critiques through (**Rebuttal**).

If this feels too long to read, below is a ‘split by major arguments’ version.

Agent + Main Action + Object + <below subarguments if used>

Context

- "Building on (Historical Context) while assuming that (Assumption), conducted under (Condition), and grounded in (Theoretical Framework). This study differentiates itself by (Differentiation)."

Purpose:

- "The goal of this work is to (Goal/Aim), which is tested by hypothesizing that (Hypothesis)."

Method:

- "The Method was conducted using (Methodology), which includes steps such as (Algorithm/Process Description) and is evaluated against (Benchmarking). The implementation was carried out using (Implementation), refined through (Optimization), and functions as described in (Mechanism)."

Results (Evaluation included):

- "The Results show (Observation), supported by (Evidence) and categorized as (Classification/Categorization). This work introduces (Novelty) and outperforms others in (Comparison). The findings are validated through (Validation) and measured using (Metrics), maintaining performance at scale through (Scalability). The results highlight relationships such as (Causation) or (Correlation), while challenges like (Limitation) and errors such as (Error/Artifact) are acknowledged within the findings."

Analysis:

- "The observed outcomes or actions are explained by (Explanation). "

Ethical Considerations:

- "This study raises concerns about (Ethical Risks) but ensures compliance through (Ethical Justification). The system benefits (Stakeholders/Beneficiaries)."

Implications:

- "The implications of this work suggest (Implication), though risks such as (Risk/Uncertainty) remain. Future work will address (Future Work/Questions)."

Contradictions to Knowledge:

- "Contrary to prior studies, this work challenges (Contradiction) and considers alternative perspectives like (Alternative/Counterargument), which are addressed through (Rebuttal)."

(for Publication)
Argument Extraction Codebook

Basic Definition (improved)

Entities: An entity is an object or set of objects in the world. Entities may be referenced in a text by their name, indicated by a common noun or noun phrase, or represented by a pronoun. (ACE 2005)

Mention: A mention is a reference to an **entity**. (ACE 2005)

Event: is a specific occurrence or happening described in text, typically involving actions, changes, or states that unfold over time. In the scientific publication, time is not usually used as a measurement of states, instead we use logical structure as guideline. Events are often annotated with key elements that provide context, such as proposal of a method, explanation of intuition and result of the method. (ACE 2005 + Modification)

Event Trigger (Main Action): The word or phrase in the text that indicates the occurrence of the event (e.g., a verb like "launched" or a noun like "discovery"). In our codebook, a similar concept **"Main Action"** is used instead of **Trigger**. **Main Action** are "Auxiliary-Compound Verb" ("am, is, are", verbs and verb phrase) that indicates the occurrence of the event. In scientific publications, authors follow a logical structured writing, which emphasis the actions they took to report new findings. In this scenario, focusing on **Main Action** which only includes verbs can better indicate the "happening" of event. Further, **Main Action** is more flexible and generalized on unseen text, while **Trigger** as defined in ACE2005 contains limited area and use case, which does not fit the fast-developing scientific domains where unseen knowledge is very frequent. (most importantly, the **Main Action** fit human reading intuition better, as it allows for diverse ground truth based on different perspective) (ACE 2005 + Modification)

Event Participants: The entities or participants involved in the event, categorized by their roles (e.g., agent, object). (ACE 2005)

Event Attributes: Additional information about the event. In our codebook's domain, this usually refers to additional scientific information such as Explanation, Results, Implication. (ACE 2005 + Modification)

Event Type: Four types of **events** are defined based on the general structure of scientific abstracts. "Background". "Methods"," Results"," Implication" (ACE 2005's idea in our domain)

Event Mention: An Event Mention is a reference to an **event**. (ACE 2005)

Arguments: Combination of **Event Participants and Event Attributes. (more detailed explanation, examples)** (ACE2005)

1. Argument

Each argument extracted should be categorized into one of the following 10 categories.

1.1 Context

- **Definition**: Provides foundational or situational information.
- **Subarguments**:
    - **Historical Context (check with Professor if this format is good)**
      Definition: Background from prior work or historical developments.
      Rationale: Historical Context provides foundational information by connecting the current study to prior developments, situating it within the broader **context** and framing its contributions and relevance**.**
      Indicator word: "builds on," "based on," "pioneered by," "founded on," "inspired by earlier work," "influenced by foundational methods."
      Examples:
      (NLP): "Builds on Word2Vec by Mikolov et al. (2013), a foundational method for word representations in NLP, we propose our method." **(Todo: indicator word color, then also highlight what to annotate)**
      (HCI): "This based on Fitts' Law (1954), a key model for predicting human movement in interaction tasks."
      (Health Informatics): "Our survey is designed by inspecting early electronic health records (EHRs) systems, pioneered in the 1960s."

    - **Assumption:**
      Definition: Underlying ideas accepted as true without direct proof.
      Rationale: Assumption identifies pre-existing conditions or untested ideas that the study relies upon, providing critical **context** for understanding the scope and setup of the research.
      Indicator Words: "Assumes that", "presumes," "is based on the premise," "underlying assumption," "implicitly assumes," "accepted without validation."
      Examples:
      (NLP): "The model assumes independence between tokens in sequence prediction tasks."
      (HCI): "The study presumes that users have prior experience with touch-screen devices."
      (Health Informatics): "The system relies on the assumption that patients accurately report symptoms in digital surveys."

    (merged below)
    - **Condition**:
      Definition: Specific limits or situational factors present during the study that influence results and define where findings apply.
      Rationale: **Condition** clarifies the specific situational scope or constraints that define the study's results, ensuring the findings are interpreted within the appropriate **context**.
      Indicator Words: "Only applies to," "limited to," "restricted to," "valid for,"

"relevant for," "conducted under," "performed in," "with the following conditions," "in a controlled environment," "under standard conditions."
Examples:
(NLP): "The evaluation was conducted on text data limited to publicly available news articles."
(HCI): "Findings are valid for desktop interfaces tested in environments with consistent lighting conditions."
(Health Informatics): "The study is restricted to patient records collected from urban hospitals over a six-month period."

- **Theoretical Framework**: (from analysis)
  Definition: Models underpinning the study.
  Rationale: **Theoretical Framework** links the study's methods or interpretations to established models or theories, ensuring that the **context** is known to readers.
  Indicator Words: "Based on," "grounded in," "applies the theory," "underpinned by," "draws on."
  Examples:
  (NLP): "This work is based on the Transformer model, which uses self-attention for sequence processing."
  (HCI): "The study applies Norman's theory of affordances to evaluate interface usability."
  (Health Informatics): "The intervention is grounded in the Health Belief Model, which predicts behavior based on perceived risks and benefits."

- **Differentiation**: (from analysis)
  Definition: Features that make this work stand out compared to others.
  Rationale: **Differentiation** provides a comparative **context** of the study's unique methods or approaches, explaining how and why they stand out or improve upon existing work.
  Indicator Words: "Unlike previous work," "stands out because," "differentiates by," "improves on," "sets itself apart."
  Examples:
  (NLP): "Unlike previous models, this approach handles multilingual text without language-specific preprocessing."
  (HCI): "This study focuses on older adults, a demographic often overlooked in interface design research."
  (Health Informatics): "Our system integrates both real-time and historical data, unlike prior models that use only static records."

1.2 Purpose (there is a relationship between purpose and method, find a way to link them in the reconstruction)

- **Definition**: Defines the purpose or aim.
- **Subarguments**:

- **Goal/Aim:**
  Definition: The main purpose or objective of the study.
  Rationale: **Goal/Aim** describes the overarching intention or end result that the study seeks to achieve, providing clear **purposes** for the research by outlining what it aims to accomplish.
  Indicator Words: "Aims to," "seeks to," "goal is," "objective is," "intends to," "purpose of the study."
  Examples:
  (NLP): "The study aims to improve sentiment classification for low-resource languages."
  (HCI): "The goal is to enhance user engagement through adaptive interface designs."
  (Health Informatics): "The aim is to evaluate the effectiveness of AI in early disease detection."

- **Hypothesis**:
  Definition: Statements proposed to explain a phenomenon, tested in the study.
  Rationale: **Hypothesis** specifies testable predictions that guide the study's methods, offering concrete **purposes** by explaining what phenomena the study investigates and seeks to validate.
  Indicator Words: "Proposes that," "tests whether," "is expected to," "hypothesizes that," "is assumed to result in," "predicts that."
  Examples:
  (NLP): "Contextual embeddings are expected to improve entity recognition accuracy."
  (HCI): "It is proposed that personalized feedback increases task completion rates."
  (Health Informatics): "This study tests whether mobile health interventions reduce patient anxiety levels."

1.3 Method

- **Definition**: Techniques, tools, or frameworks used.
- **Subarguments**:
  - **Methodology**:
    Definition: General description of methods used.
    Rationale: **Methodology** outlines the overall approach or framework used in the study, providing a high-level description of the **methods** applied to achieve the study's goals.
    Indicator Words: "Was conducted using," "employed," "utilized," "was analyzed with," "applied methods include," "Updates based on," "refined iteratively," "adjusted using feedback," "loop includes."
    Examples:
    (NLP): "Transformer-based models were fine-tuned on annotated datasets."
    (HCI): "Usability testing was conducted with think-aloud protocols."

(Health Informatics): "Patient surveys were analyzed using mixed-methods approaches."

- **Subsequent Methodology**:
  Definition: Additional or follow-up techniques and procedures applied after the primary methodology.
  Rationale: **Subsequent methodologies** are critical for extending, validating, or refining initial findings. They establish a logical progression in the **method** and often bridge initial approaches with additional results or validation steps**.**
  Indicator Words: "Following this," "subsequent steps include," "in the next phase," "was further refined by," "extended through."
  Examples:
  (NLP): "Following this, subsequent steps included fine-tuning the model on domain-specific datasets to improve accuracy further."
  (HCI): "Subsequent methods involved iterative prototyping and user feedback collection to refine the interface."
  (Health Informatics): "In the next phase, machine learning models were retrained with additional features to enhance diagnostic precision."

- **Algorithm/Process Description**:
  Definition: Step-by-step technical details.
  Rationale: **Algorithm/Process Description** provides a detailed breakdown of the procedural steps or workflows involved, explaining how the **methods** were implemented technically..
  Indicator Words: "Preprocesses by," "includes steps such as," "consists of," "process involves," "workflow includes."
  Examples:
  (NLP): "The system preprocesses text by tokenizing, normalizing, and embedding inputs before classification."
  (HCI): "The interface design process included wireframing, prototyping, and iterative user testing."
  (Health Informatics): "The prediction pipeline involves feature extraction, model training, and validation on patient datasets."

(todo: Benchmarking maybe in analysis?)

- **Benchmarking**:
  Definition: Comparisons against standards or other methods.
  Rationale: **Benchmarking** evaluates the effectiveness of the study's methods by comparing results against established baselines or alternative approaches, demonstrating the **methods**' relative performance.
  Indicator Words: "Evaluated against," "compared to," "benchmark results include," "was measured relative to," "tested alongside."
  Examples:
  (NLP): "The model was evaluated against BERT and GPT baselines on standard

benchmarks."
(HCI): "User performance was compared to established metrics from prior usability studies."
(Health Informatics): "Diagnostic accuracy was benchmarked against results from clinical experts."

- **Implementation**:
  Definition: Practical realization of theories or tools.
  Rationale: **Implementation** describes how the theories, methods, or frameworks were practically realized, providing concrete details on the **methods**' real-world execution.
  Indicator Words: "Implemented using," "was developed as," "deployed with," "realized in," "translated into."
  Examples:
  (NLP): "The model was implemented using PyTorch and trained on a distributed GPU cluster."
  (HCI): "A prototype application was developed using React and deployed for user testing."
  (Health Informatics): "The intervention was implemented as a mobile app integrated with existing EHR systems."

- **Optimization**:
  Definition: Refinements for better outcomes.
  Rationale: **Optimization** focuses on adjustments or refinements made to improve the efficiency, accuracy, or performance of the **methods** used in the study.
  Indicator Words: "Optimized by," "tuned using," "refined to," "adjusted for," "improved to maximize."
  Examples:
  (NLP): "Hyperparameters were tuned using grid search to maximize classification accuracy."
  (HCI): "Interaction flows were streamlined to reduce task completion time."
  (Health Informatics): "The algorithm was optimized for faster processing of large-scale patient datasets."

- **Mechanism**:
  Definition: The step-by-step process explaining how something works.
  Rationale: **Mechanism** provides a detailed breakdown of how methods or processes function, offering a systematic **methods** of why specific approaches produce the observed outcomes.
  Indicator Words: "Works by," "operates through," "step-by-step process," "mechanism involves," "functioning of."
  Examples:
  (NLP): "The attention mechanism assigns weights to tokens, prioritizing more

relevant context for predictions."
(HCI): "The recommendation system adjusts displayed content based on real-time user interactions."
(Health Informatics): "The wearable device measures heart rate variability to detect early signs of stress."

- **Validation**:
  Definition: Checking to make sure the results are correct and reliable.
  Rationale: **Validation** ensures the correctness and reliability of the research by using verification **methods** to confirm their accuracy and robustness.
  Indicator Words: "Validated through," "confirmed by," "cross-checked with," "verified using."
  Examples:
  (NLP): "The model's predictions were validated using cross-validation on three datasets."
  (HCI): "Usability findings were confirmed through follow-up user testing sessions."
  (Health Informatics): "The algorithm's outputs were validated against expert-annotated clinical data."

1.4 Results (Evaluation included)

- **Definition**: Observations or outputs.
- **Subarguments**:
  - **Observation**:
    Definition: Important findings or patterns noticed during the study.
    Rationale: **Observation** identifies key patterns or findings that emerge from the study's analysis, providing direct **results** that highlight what was discovered.
    Indicator Words: "Observed that," "noticed a trend where," "found that," "results showed."
    Examples:
    (NLP): "The model struggled with rare words, showing lower accuracy for low-frequency tokens."
    (HCI): "Users preferred interfaces with fewer menu options, reporting higher satisfaction scores."
    (Health Informatics): "We found that patients responded more quickly to reminders sent in the morning compared to the evening."

  - **Evidence/Support**:
    Definition: Data or results used to back up conclusions.
    Rationale: **Evidence/Support** provides concrete data that substantiates the study's findings, presenting measurable **results** to validate conclusions.
    Indicator Words: "Achieved," "supported by," "data shows," "results confirm."
    Examples:
    (NLP): "The model achieved 92% accuracy on the test set, outperforming all baselines."

(HCI): "80% of participants completed the task faster using the new interface design."
(Health Informatics): "Survey data showed a 30% reduction in reported symptoms after the intervention."

- **Classification/Categorization**:
  Definition: Organizing data into specific groups or labels.
  Rationale: **Classification/Categorization** organizes the study's findings into meaningful groups, providing structured **results** that simplify analysis and interpretation.
  Indicator Words: "Organized into," "grouped by," "categorized as," "labeled with."
  Examples:
  (NLP): "Text data was categorized into sentiment labels: positive, negative, or neutral."
  (HCI): "Users were grouped based on their interaction styles: novice, intermediate, or expert."
  (Health Informatics): "Patients were classified by risk level: low, moderate, or high."

(todo: have better examples)

- **Novelty**:
  Definition: New and original contributions or discoveries.
  Rationale: **Novelty** highlights unique findings or innovations introduced by the study, presenting **results** that demonstrate its originality and contribution to the field.
  Indicator Words: "Introduces," "presents a new," "first to," "novel contribution."
  Examples:
  (NLP): "This study introduces a new algorithm for real-time translation of code-mixed languages."
  (HCI): "We present a novel interface that adapts dynamically based on user emotions."
  (Health Informatics): "Our work is the first to integrate wearable device data into predictive health models."

Evaluation: evidence and supportive information of Results (Combine these because they are often inseparable)

- **Comparison**:
  Definition: Comparing of findings or approaches.
  Rationale: **Comparison** evaluates the study's **results** relative to other methods or baselines, highlighting its performance, advantages, or differences.
  Indicator Words: "Outperforms," "compared to," "similar to," "results indicate

better than."
Examples:
(NLP): "Our model outperforms GPT-3 in accuracy but requires less computational resources."
(HCI): "This design resulted in faster task completion compared to traditional menu-based interfaces."
(Health Informatics): "The proposed algorithm shows similar precision to expert diagnoses but operates significantly faster."

- **Metrics**:
  Definition: Measures used to evaluate the performance or quality of a system or method.
  Rationale: Metrics quantify the **results** by providing measurable criteria that assess the performance, quality, or effectiveness of the study's methods.
  Indicator Words: "Measured by," "evaluated using," "performance was assessed with," "scores indicate."
  Examples:
  (NLP): "The model's performance was evaluated using accuracy, F1 score, and BLEU for translation tasks."
  (HCI): "User satisfaction was measured through SUS (System Usability Scale) scores and task completion time."
  (Health Informatics): "The system's effectiveness was assessed using sensitivity, specificity, and patient adherence rates."

- **Scalability**:
  Definition: The ability of a system to handle larger or more complex tasks effectively.
  Rationale: **Scalability** evaluates how well the **results** hold up as the system is scaled to handle larger datasets, tasks, or environments.
  Indicator Words: "Scales to," "adapts seamlessly," "handles large-scale," "maintains performance at scale."
  Examples:
  (NLP): "The architecture scales efficiently to datasets with billions of sentences."
  (HCI): "The interface adapts seamlessly to multi-device environments, including tablets and smartphones."
  (Health Informatics): "The platform supports integration with healthcare systems managing millions of patient records."

Was in Analysis

- **Causation**:
  Definition: Direct relationships where one thing causes another.
  Rationale: **Causation** identifies cause-effect relationships observed in the study, showing how changes in one variable directly influence **results**.
  Indicator Words: "Results show that," "caused by," "led to," "directly influenced

by."
Examples:
(NLP): "Increasing training data size directly improves the model's accuracy."
(HCI): "Reducing menu complexity led to faster task completion times."
(Health Informatics): "Timely medication reminders significantly increased patient adherence rates."

- **Correlation**:
  Definition: Relationships observed between variables, without proving one causes the other.
  Rationale: **Correlation** highlights relationships or associations observed in the **results**, without establishing direct causality.
  Indicator Words: "Associated with," "linked to," "correlates with," "relationship between."
  Examples:
  (NLP): "Word frequency correlates with the model's confidence in its predictions."
  (HCI): "Longer session times are associated with higher user satisfaction ratings."
  (Health Informatics): "Higher social media activity is linked to increased patient engagement with health interventions."

- **Disagreement**:
  Definition: Unexpected differences between observed and expected results.
  Rationale: **Disagreement** identifies inconsistencies between predictions and actual **Results**.
  Indicator Words: "Contrary to expectations," "unexpectedly," "differences between," "observed versus expected," "discrepancies in."
  Examples:
  (NLP): "The model performed worse on shorter sentences than anticipated."
  (HCI): "Contrary to expectations, users preferred the non-adaptive interface in time-sensitive tasks."
  (Health Informatics): "Predicted recovery rates were significantly lower than clinical observations."

1.5 Analysis (Why using some method)

- **Definition**: Interpretation or theoretical explanation of **Methods and Results**.
- **Subarguments**:
  - **Explanation**:
    Definition: Reasons given to clarify why something happens.
    Rationale: **Explanation** provides interpretations of the study's observations by clarifying why certain methods or phenomena occur, contributing to the overall **analysis** of the study's choices and findings.
    Indicator Words: "Because," "due to," "as a result of," "this happens when,"

"explained by."
Examples:
(NLP): "The model struggles with rare words because the training data lacks sufficient examples."
(HCI): "Users prefer larger buttons due to improved visibility and easier interaction."
(Health Informatics): "Patients responded more quickly to shorter reminders because they require less cognitive effort."

- **Examples**:
  Definition: Specific instances or scenarios used to illustrate findings, phenomena, or explanations.
  Rationale: **Examples** help contextualize and clarify abstract findings or explanations, making them more tangible and relatable. They enhance the **analysis** by providing real-world relevance or detailed case scenarios.
  Indicator Words: "For instance," "such as," "an example is," "illustrated by," "demonstrated through."
  Examples:
  (NLP): "An example of this is the model's performance improvement on domain-specific text, such as medical records."
  (HCI): "For instance, older adults demonstrated higher interaction satisfaction when larger buttons were introduced."
  (Health Informatics): "This can be seen in cases where early intervention prevented adverse health outcomes, such as reduced hospital readmissions."

- **Elaborations**:
  Definition: Detailed expansions or additional explanations that clarify findings, methods, or theoretical insights.
  Rationale: **Elaborations** provide deeper insights or further interpret findings and observations, enhancing understanding. They enrich the **analysis** by breaking down complex concepts into more digestible details or exploring implications in depth.
  Indicator Words: "This means that," "in more detail," "can be expanded as," "to elaborate further".
  Examples:
  (NLP): "This means that increasing the training corpus size directly affects performance for low-resource languages."
  (HCI): "To elaborate further, user satisfaction improves when interfaces are tailored to individual preferences."
  (Health Informatics): "This can be expanded as a multi-stage process where wearable devices contribute both real-time and historical data for predictions."

1.6 Challenge (todo: maybe also in result?)

- **Definition**: Constraints or weaknesses.

- **Subarguments**:
    - **Limitation**:
      Definition: Factors that may affect the reliability or validity of the results.
      Rationale: **Limitation** highlights constraints that influence the reliability or applicability of the study's findings, making it a critical part of the overall **Challenge** by identifying weaknesses that frame the study's scope.
      Indicator Words: "Limited by," "constrained by," "restricted to," "affected by," "scope includes."
      Examples:
      (NLP): "The model's performance is limited by the lack of diverse training data."
      (HCI): "Findings are constrained by the use of a controlled lab environment, which may not reflect real-world conditions."
      (Health Informatics): "The analysis is limited to data collected from a single healthcare provider."

    - **Error/Artifact**:
      Definition: Issues or anomalies that may affect the accuracy or interpretation of results.
      Rationale: **Error/Artifact** addresses inaccuracies or irregularities in the data or methods, contributing to the **Challenge** by identifying potential distortions that impact result interpretation.
      Indicator Words: "May include errors," "affected by artifacts," "issues in," "noise in data," "anomalies from."
      Examples:
      (NLP): "The model occasionally misclassifies due to noise in the training data."
      (HCI): "User testing results may include artifacts from inconsistent internet speeds during sessions."
      (Health Informatics): "Measurement errors in wearable devices could impact the accuracy of health predictions."

1.7 Ethical

- **Definition**: Ethical concerns, implications, and justifications related to the study or system.
- **Subarguments**:
    - **Ethical Risks**:
      Definition: Potential moral or societal risks arising from the study or system.
      Rationale: **Ethical Risks** identify specific negative ethical implications, such as bias, fairness, privacy violations, or societal harm, which need to be addressed as part of **Ethical Considerations**.
      Indicator Words: "Raises concerns about," "risks of," "ethical implications," "must ensure," "could perpetuate," "potential harm."
      Examples:
      (NLP): "Deploying biased language models risks reinforcing societal stereotypes

and marginalizing vulnerable groups."
(HCI): "Persuasive interface designs may exploit users' cognitive biases without their informed consent."
(Health Informatics): "Opaque AI-driven diagnoses can undermine patient trust and lead to inequitable healthcare outcomes."

- **Ethical Justification**:
  Definition: Reasons given to support ethical choices in a study or system.
  Rationale: **Ethical Justification** explains the steps taken to mitigate ethical risks and ensure the study aligns with accepted moral principles, contributing to the broader **Ethical Considerations** argument.
  Indicator Words: "Ensures privacy by," "justified by," "prioritizes fairness," "adheres to ethical guidelines," "promotes inclusivity," "balances utility and ethics."
  Examples:
  (NLP): "Training data was anonymized to protect user privacy while ensuring model utility."
  (HCI): "The system prioritizes accessibility features to promote inclusivity for users with disabilities."
  (Health Informatics): "Patient consent was obtained before using data to ensure adherence to ethical guidelines."

- **Stakeholders/Beneficiaries**:
  Definition: Groups or individuals affected by or benefiting from the study or system.
  Rationale: **Stakeholders/Beneficiaries** identify the groups impacted by the study or system, ensuring that **Ethical Considerations** account for fairness in distributing benefits and acknowledging societal impacts.
  Indicator Words: "Benefits include," "stakeholders impacted are," "beneficiaries gain from," "affects groups such as," "end-users benefit from."
  Examples:
  (NLP): "Developers of multilingual applications benefit from the improved translation model."
  (HCI): "End-users of e-learning platforms gain from the enhanced interface usability."
  (Health Informatics): "Patients with chronic conditions benefit from real-time monitoring and predictive alerts."

1.8 Implications

- **Definition**: Broader applicability or significance and future research suggestion.
- **Subarguments**:
  - **Implication**:
    Definition: The broader significance or potential impact of the findings.
    Rationale: **Implication** highlights the broader applicability or real-world impact of

the study's findings, demonstrating the relevance and significance of the research within the **Implications** argument.
Indicator Words: "Suggests that," "could improve," "may lead to," "potential impact includes," "significant for," "enhances."
Examples:
(NLP): "Improved sentiment analysis can enhance content moderation on social media platforms."
(HCI): "Findings suggest that adaptive interfaces could significantly improve productivity in workplace applications."
(Health Informatics): "Real-time health monitoring may reduce hospital readmission rates through early intervention."

- **Risk/Uncertainty**:
  Definition: Potential challenges or unknown factors that could affect outcomes.
  Rationale: **Risk/Uncertainty** identifies potential limitations or unknowns regarding the applicability or success of the study's findings, ensuring that the **Implications** account for real-world challenges and variability.
  Indicator Words: "May fail when," "uncertainty about," "risks include," "unknown factors," "could be affected by," "introduces challenges."
  Examples:
  (NLP): "The model may fail when applied to languages not represented in the training data."
  (HCI): "There is uncertainty about how users will adapt to the new interface in high-stress environments."
  (Health Informatics): "Reliance on wearable devices introduces the risk of inaccurate health data due to hardware malfunctions."

- **Future Work/Questions**:
  Definition: Suggested directions for further research or development.
  Rationale: **Future Work/Questions** proposes extensions, refinements, or new research directions, showing how the study's findings can inspire continued exploration under the **Implications** argument.
  Indicator Words: "Future work includes," "further studies are needed," "could explore," "suggested directions include," "future research could."
  Examples:
  (NLP): "Future work could explore adapting the model for low-resource languages."
  (HCI): "Further studies are needed to evaluate interface performance in real-world settings."
  (Health Informatics): "Future research could investigate the integration of genomic data into predictive health models."

1.9 Contradictions to (current) Knowledge (in the analysis, or result, not frequent so merge it)

- **Definition**: Points challenging existing knowledge.
- **Subarguments**:
    - **Contradiction**:
      Definition: Findings that conflict with or challenge prior research.
      Rationale: **Contradiction** highlights results that directly challenge or deviate from existing knowledge, contributing to the **Contradictions to Knowledge** argument by questioning established assumptions or findings.
      Indicator Words: "Unlike prior studies," "contrary to previous research," "our findings challenge," "in conflict with," "disagrees with."
      Examples:
      (NLP): "Unlike prior studies, our results show that pretraining on domain-specific data is not always beneficial."
      (HCI): "Contrary to previous research, we found that users preferred static menus over adaptive ones."
      (Health Informatics): "Our findings contradict earlier claims that wearable devices always improve patient adherence."

    - **Alternative/Counterargument**:
      Definition: Differing or opposing perspectives on an issue.
      Rationale: **Alternative/Counterargument** introduces opposing viewpoints or perspectives that challenge the study's claims, enhancing the **Contradictions to Knowledge** argument by presenting alternative explanations or frameworks.
      Indicator Words: "An alternative view suggests," "some argue that," "a counter perspective is," "opposing viewpoints include," "it has been proposed that."
      Examples:
      (NLP): "An alternative view suggests that rule-based systems are more interpretable than neural models for certain tasks."
      (HCI): "Some argue that traditional desktop interfaces remain more efficient than mobile-first designs for complex workflows."
      (Health Informatics): "A counterargument is that patient-reported data might be less reliable than clinical observations."

    - **Rebuttal**:
      Definition: Responses addressing critiques or opposing views.
      Rationale: **Rebuttal** provides responses or justifications to counter opposing views, strengthening the **Contradictions to Knowledge** argument by defending the study's claims against alternative perspectives.
      Indicator Words: "While it is true that," "although some argue," "despite concerns," "however," "nonetheless."
      Examples:
      (NLP): "While rule-based systems are interpretable, they lack the flexibility and scalability of neural models."
      (HCI): "Although desktop interfaces may be efficient, mobile-first designs better meet the needs of on-the-go users."

(Health Informatics): "Despite concerns about reliability, patient-reported data provides valuable insights when combined with clinical records."

2. Annotation Guidelines

1. First read the paragraph and summarize an **Event mention (which can be seen as title)** that best describes the paragraph.
2. With the help from **Event mention**, find an "Auxiliary-Compound Verb" ("am, is, are" and verbs) from the paragraph that best represents the **Event mention**. This is called the **"Main Action"** and it has to be an exact quote in the paragraph.
3. Then in this paragraph, find a "main sentence" using the **"Main Action"**, and in most of the cases, "main sentence" is the full sentence that contains the **"Main Action"**. All arguments extracted should be describing this "main sentence" in different ways. Note that there could be other same **"Main Action"** words (as in the English aspect) used in different sentences, you should treat them differently and they should not be the **"Main Action"**. E.g.,
4. There are two additional arguments to find first. The doer of the main action called **"Agent"** and receiver of the action called **"Object". In the case of passive tense, e.g. A is proposed by B; you should extract Agent: "B", Main Action: "propose", Object: "A".**
5. Use the **"Main action"** to extract other **arguments**. Then use subarguments above to fit the extracted arguments and find its argument argument. Don't infer any information, use only the exact same quote from the given paragraphs. You don't need to fill all 10 categories for each paragraph, but you are encouraged to cover as much information as possible.
6. You should always try to extract compound nouns only, unless one entire sentence is describing one argument, and the sentence is inseparable in meaning. **When encountered abbreviation, e.g. "Chain-of-Thought (CoT) improve …...", you should use both original term and abbreviation: "Chain-of-Thought (CoT)" as the argument.**
7. Following from 6, argument boundaries should not include pronouns unless essential. E.g. "To explore the impact on creators, we present an interview based study of journalists and political commentators on YouTube." As instructed by 6 and 7, choose only an interview based study.
8. In the end, you should annotate **Event mention, Main action, 9+2 Arguments** (if explicit)

9. Diverse annotations are encouraged! If you check the reconstructed text and find multiple ways of annotating one paragraph, please take note of the paper code, text, and what you think might also be appropriate annotations!

10. Different versions will be kept. **(all are ground truth?) (later use (majority vote) high agreement data 80% to be no dispute.)**

Notes on disagreements: (need a few examples)

1. We should focus on what the author is trying to focus on. E.g., in structure like “Context + However + challenge” our main action should be from “challenge” part, because author focus more on challenge when writing the paper.

2. Boundary specification of object: Object will follow this structure: {“**Base object**”: “”, “**Modifier**”: “”, “**Attached Object**”: “”, “**Modifier"**: “”} to solve the issue.
Definitions:
Base object: The primary object without any descriptor.
Attached object: The secondary object that is directly related to the base object (also without any descriptor).
Modifier: The descriptor of base or attached object.
Here is an example: Understanding the full spectrum of uptake factors is essential to identify ways in which policy makers and providers can facilitate the adoption of effective digital therapeutics within a health system, as well as the steps developers can take to assist in the deployment of products.

Main Action: “is essential to identify”
Object: {“**Base object**”: “ways”, “**Modifier**”: “in which policy makers and providers can facilitate the adoption of effective digital therapeutics within a health system”, “**Attached Object**”: “as well as the steps”, “**Modifier"**: “developers can take to assist in the deployment of products.”}

Structure like “Map A into B” can be also solved using this formulation of object, where A is **base object**, and into B as the **attached object. Attached object should include the proposition before it, to indicate the relation between it and the base object.**

3. We don’t have an explicit **Object** sometimes, and in that case, we leave **Object** empty:
e.g.,
indicate that; (Verb + that + clause)
conclude by doing (verb + prop +doing)

4. Main action should include every word if it is a Predicative Infinitive Constructions. E.g., be the first to do”, “have the potential to do”, “be essential to do” (from above)

5. Overlapping issues:

- Method v.s. Context, e.g., with …… we obtain, using …… we obtain. **Method** tends to discuss more detailed design, while **Context** simply mention the name of a method.

- Challenge v.s. Result and Analysis: **Challenge** can sometimes also fit under **Results or Analysis (examples)**, but we enforce a **more detailed annotation when possible**.
  E.g.,
  Findings reveal the high-stakes articulation labor demanded of workers to be recognized by these systems, including maintaining multiple mobile devices, repeatedly uploading requisite images, spending time and resources visiting customer-service centers, and making physical changes to their bodies and environments.
  Highlighted text should be **challenge** rather than **example** (in analysis), because challenge is **more fine-grained** definition.

3. Annotation Example

Below is an example that follows the above guideline.

We propose a genetic algorithm (GA) based method for modifying n-best lists produced by a machine translation (MT) system. Our method offers an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics

1. The **Event mention** of this is “GA based method proposal”
2. The **main action** is: “propose”.
3. The main sentence is: We propose a genetic algorithm (GA) based method for modifying n-best lists produced by a machine translation (MT) system.
4. Find “**agent**” to be “We”, “**object**” to be “a genetic algorithm (GA) based method”
5. “for modifying n-best lists produced by a machine translation (MT) system” can be seen as “Goal/Aim” in the “**Purpose**” argument.

   “an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics” is the “Novelty” in the “**result**” argument.

6. Why “for modifying n-best lists produced by a machine translation (MT) system” shouldn’t be separated?

   This is because, the “produced by a machine translation (MT) system” is a strong description of “n-best lists”, and splitting them is wrong, because “produced by a machine translation (MT) system” is not describing the “main action” (which is connected to “GA based method”).

7. A good way to check is by reconstruction and check if the meaning is consistent.

4. Reconstruction Template (self-validate) (pre-validate quality) (multiple mention template sets)

(we extract and reconstruct our ground truth as list, and compare with the list from annotators)
Agent + Main Action + Object + In the **context** of _____ + to \ for \ because ____ (**purpose**) + using ____(**method**) and (todo, delete _____ **evaluation** metric), and the **result** is ___ + and all of the above **implies** ____. This is because ___(**analysis**). This **contradicts** ___. This has some known **limitations**, such as ___. Additionally, **ethical considerations** like _____ are also a part of consideration.

Note that the above template is describing “Main Actions”; All above arguments should be expanded from the “Main sentence”.

Example reconstructed:

We + propose + a genetic algorithm (GA) based method + for “modifying n-best lists produced by a machine translation (MT) system”, and the result is “an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics”.

We compare the meaning of the reconstructed example and found it to be consistent with the original full paragraph.

(for student)

Argument Extraction Codebook

Basic Definition (improve this)

Event: is a specific occurrence or happening described in text, typically involving actions, changes, or states that unfold over time. Events are often annotated with key elements that provide context, such as:

Event Trigger (Main Action): The word or phrase in the text that indicates the occurrence of the event (e.g., a verb like "launched" or a noun like "discovery"). In our codebook, a similar concept **"Main Action"** is used instead of **Trigger**. **Main Action** are "Auxiliary-Compound Verb" ("am, is, are", verbs and verb phrase)

Event Participants: The entities or participants involved in the event, categorized by their roles (e.g., agent, object).

Event Attributes: Additional information about the event. In our codebook's domain, this usually refers to additional scientific information such as Explanation, Results, Implication.

Event Type: Four types of **events** are defined based on the general structure of scientific abstracts. "Background". "Methods","Results","Implication"

Event Mention: A name that best summarizes the **event**.

Arguments: Combination of **Event Participants and Event Attributes.**

1. Argument Argument

Each argument extracted should be categorized into one of the following 10 categories.

1.1 Context

- **Definition**: Provides foundational or situational information..
- **Subarguments**:
    - **Historical Context:** Background from prior work or historical developments. **(highlight the subarguments)**

(TODO: rationale, indicator word)
(NLP) E.g., "This builds on Word2Vec by Mikolov et al. (2013), a foundational method for word representations in NLP."
(HCI) E.g., "This builds on Fitts' Law (1954), a key model for predicting human movement in interaction tasks."
(Health Informatics) E.g., "This builds on early electronic health records (EHRs) systems, pioneered in the 1960s."

- Assumption: Underlying ideas accepted as true without direct proof.
  (NLP) E.g., "The model assumes independence between tokens in sequence prediction tasks."
  (HCI) E.g., "The study assumes that users have prior experience with touch-screen devices."
  (Health Informatics) E.g., "The system assumes patients accurately report symptoms in digital surveys."

- Boundary Condition: Specific limits or contexts where findings apply.
  (NLP) E.g., "The results are valid only for datasets with English-language text."
  (HCI) E.g., "Findings apply to tasks performed on mobile devices with screens smaller than 7 inches."
  (Health Informatics) E.g., "The analysis is limited to data from patients aged 18 to 65 years."

- Normal Condition: Situational factors present during the study that influence results.
  (NLP) E.g., "The experiments were conducted using GPU-accelerated training environments."
  (HCI) E.g., "User testing was performed in a controlled lab environment with stable internet connectivity."
  (Health Informatics) E.g., "The study was conducted during regular outpatient clinic hours."

1.2 Purpose

- **Definition**: Defines the purpose or aim.
- **Subarguments**:
  - Goal/Aim: The main purpose or objective of the study.
    (NLP) E.g., "The study aims to improve sentiment classification for low-resource languages."
    (HCI) E.g., "The goal is to enhance user engagement through adaptive interface designs."
    (Health Informatics) E.g., "The aim is to evaluate the effectiveness of AI in early disease detection."

- Hypothesis: Statements proposed to explain a phenomenon, tested in the study.
  (NLP) E.g., "Contextual embeddings are expected to improve entity recognition accuracy."
  (HCI) E.g., "It is proposed that personalized feedback increases task completion rates."
  (Health Informatics) E.g., "This study tests whether mobile health interventions reduce patient anxiety levels."

1.3 Method

- **Definition**: Techniques, tools, or frameworks used.
- **Subarguments**:
  - Methodology: General description of methods used.
    (NLP) E.g., "Transformer-based models were fine-tuned on annotated datasets."
    (HCI) E.g., "Usability testing was conducted with think-aloud protocols."
    (Health Informatics) E.g., "Patient surveys were analyzed using mixed-methods approaches."

  - Algorithm/Process Description: Step-by-step technical details.
    (NLP) E.g., "The system preprocesses text by tokenizing, normalizing, and embedding inputs before classification."
    (HCI) E.g., "The interface design process included wireframing, prototyping, and iterative user testing."
    (Health Informatics) E.g., "The prediction pipeline involves feature extraction, model training, and validation on patient datasets."

  - Benchmarking: Comparisons against standards or other methods.
    (NLP) E.g., "The model was evaluated against BERT and GPT baselines on standard benchmarks."
    (HCI) E.g., "User performance was compared to established metrics from prior usability studies."
    (Health Informatics) E.g., "Diagnostic accuracy was benchmarked against results from clinical experts."

  - Implementation: Practical realization of theories or tools.
    (NLP) E.g., "The model was implemented using PyTorch and trained on a distributed GPU cluster."
    (HCI) E.g., "A prototype application was developed using React and deployed for user testing."
    (Health Informatics) E.g., "The intervention was implemented as a mobile app integrated with existing EHR systems."

  - Optimization: Refinements for better outcomes.
    (NLP) E.g., "Hyperparameters were tuned using grid search to maximize classification accuracy."

(HCI) E.g., "Interaction flows were streamlined to reduce task completion time."
(Health Informatics) E.g., "The algorithm was optimized for faster processing of large-scale patient datasets."

- Impact of Size: Examination of changes with scale.
  (NLP) E.g., "The model maintains accuracy when scaled to datasets with billions of tokens."
  (HCI) E.g., "User satisfaction decreases slightly as the number of interface options increases."
  (Health Informatics) E.g., "The system performs reliably when processing data from thousands of patients simultaneously."

1.4 Results

- **Definition**: Observations or outputs.
- **Subarguments**:
  - Observation: Important findings or patterns noticed during the study.
    (NLP) E.g., "The model struggled with rare words, showing lower accuracy for low-frequency tokens."
    (HCI) E.g., "Users preferred interfaces with fewer menu options, reporting higher satisfaction scores."
    (Health Informatics) E.g., "Patients responded more quickly to reminders sent in the morning compared to the evening."

  - Evidence/Support: Data or results used to back up conclusions.
    (NLP) E.g., "The model achieved 92% accuracy on the test set, outperforming all baselines."
    (HCI) E.g., "80% of participants completed the task faster using the new interface design."
    (Health Informatics) E.g., "Survey data showed a 30% reduction in reported symptoms after the intervention."

  - Classification/Categorization: Organizing data into specific groups or labels.
    (NLP) E.g., "Text data was categorized into sentiment labels: positive, negative, or neutral."
    (HCI) E.g., "Users were grouped based on their interaction styles: novice, intermediate, or expert."
    (Health Informatics) E.g., "Patients were classified by risk level: low, moderate, or high."

  - Novelty: New and original contributions or discoveries.
    (NLP) E.g., "This study introduces a **new** algorithm for real-time translation of code-mixed languages."
    (HCI) E.g., "We present a **novel** interface that adapts dynamically based on user emotions."

(Health Informatics) E.g., "Our work is the **first** to integrate wearable device data into predictive health models."

- Feedback Loop: Repeated processes where outputs influence future actions or results.
  (NLP) E.g., "The model updates embeddings based on user corrections to improve future predictions."
  (HCI) E.g., "User behavior data is collected and used to refine interface recommendations."
  (Health Informatics) E.g., "Patient responses to alerts adjust the timing and content of subsequent notifications."

1.5 Analysis

- **Definition**: Interpretation or theoretical explanation of results.
- **Subarguments**:
  - Explanation: Reasons given to clarify why something happens.
    (NLP) E.g., "The model struggles with rare words because the training data lacks sufficient examples."
    (HCI) E.g., "Users prefer larger buttons due to improved visibility and easier interaction."
    (Health Informatics) E.g., "Patients responded more quickly to shorter reminders because they require less cognitive effort."

  - Mechanism: The step-by-step process explaining how something works.
    (NLP) E.g., "The attention mechanism assigns weights to tokens, prioritizing more relevant context for predictions."
    (HCI) E.g., "The recommendation system adjusts displayed content based on real-time user interactions."
    (Health Informatics) E.g., "The wearable device measures heart rate variability to detect early signs of stress."

  - Theoretical Framework: Models underpinning the study.
    (NLP) E.g., "This work is based on the Transformer model, which uses self-attention for sequence processing."
    (HCI) E.g., "The study applies Norman's theory of affordances to evaluate interface usability."
    (Health Informatics) E.g., "The intervention is grounded in the Health Belief Model, which predicts behavior based on perceived risks and benefits."

  - Causation: Direct relationships where one thing causes another.
    Note: Not in result? Because it focuses more on why something happened, rather than simply report results.
    (NLP) E.g., "Increasing training data size directly improves the model's accuracy."

(HCI) E.g., "Reducing menu complexity led to faster task completion times."
(Health Informatics) E.g., "Timely medication reminders significantly increased patient adherence rates."

- Correlation: Relationships observed between variables, without proving one causes the other.
  (NLP) E.g., "Word frequency correlates with the model's confidence in its predictions."
  (HCI) E.g., "Longer session times are associated with higher user satisfaction ratings."
  (Health Informatics) E.g., "Higher social media activity is linked to increased patient engagement with health interventions."

- Differentiation: Features that make this work stand out compared to others.
  (NLP) E.g., "Unlike previous models, this approach handles multilingual text without language-specific preprocessing."
  (HCI) E.g., "This study focuses on older adults, a demographic often overlooked in interface design research."
  (Health Informatics) E.g., "Our system integrates both real-time and historical data, unlike prior models that use only static records."

- Trade-off: Balancing compromises between two competing factors.
  (NLP) E.g., "The model achieves higher accuracy at the cost of increased computational time."
  (HCI) E.g., "Simplifying the interface improved usability but reduced the feature set."
  (Health Informatics) E.g., "Incorporating detailed patient data improves predictions but increases data processing time."

1.6 Evaluation

- **Definition**: Quality or relevance compared to others.
- **Subarguments**:
  - Comparison: Juxtaposition of findings or approaches.
    (NLP) E.g., "Our model outperforms GPT-3 in accuracy but requires less computational resources."
    (HCI) E.g., "This design resulted in faster task completion compared to traditional menu-based interfaces."
    (Health Informatics) E.g., "The proposed algorithm shows similar precision to expert diagnoses but operates significantly faster."

  - Validation: Checking to make sure the results are correct and reliable.
    (should be in the evaluation?)
    E.g., (NLP) E.g., "The model's predictions were validated using cross-validation on three datasets."

(HCI) E.g., "Usability findings were confirmed through follow-up user testing sessions."
(Health Informatics) E.g., "The algorithm's outputs were validated against expert-annotated clinical data."

- Metrics: Measures used to evaluate the performance or quality of a system or method.
  (NLP) E.g., "The model's performance was evaluated using accuracy, F1 score, and BLEU for translation tasks."
  (HCI) E.g., "User satisfaction was measured through SUS (System Usability Scale) scores and task completion time."
  (Health Informatics) E.g., "The system's effectiveness was assessed using sensitivity, specificity, and patient adherence rates."

- Scalability: The ability of a system to handle larger or more complex tasks effectively.
  (NLP) E.g., "The architecture scales efficiently to datasets with billions of sentences."
  (HCI) E.g., "The interface adapts seamlessly to multi-device environments, including tablets and smartphones."
  (Health Informatics) E.g., "The platform supports integration with healthcare systems managing millions of patient records."

1.7 Challenge

- **Definition**: Constraints or weaknesses.
- **Subarguments**:
  - Limitation: Factors that may affect the reliability or validity of the results.
    (NLP) E.g., "The model's performance is limited by the lack of diverse training data."
    (HCI) E.g., "Findings are constrained by the use of a controlled lab environment, which may not reflect real-world conditions."
    (Health Informatics) E.g., "The analysis is limited to data collected from a single healthcare provider."

  - Error/Artifact: Issues or anomalies that may affect the accuracy or interpretation of results.
    (NLP) E.g., "The model occasionally misclassifies due to noise in the training data."
    (HCI) E.g., "User testing results may include artifacts from inconsistent internet speeds during sessions."
    (Health Informatics) E.g., "Measurement errors in wearable devices could impact the accuracy of health predictions."

- Disagreement: Unexpected differences between observed and expected results.
  (NLP) E.g., "The model performed worse on shorter sentences than anticipated."
  (HCI) E.g., "Contrary to expectations, users preferred the non-adaptive interface in time-sensitive tasks."
  (Health Informatics) E.g., "Predicted recovery rates were significantly lower than clinical observations."

1.8 Ethical Considerations

- **Definition**: Ethical consideration and statement.
- **Subarguments**:
  - Ethical Considerations: Concerns about the moral implications of a study or system.
    Note: Why not in limitation: Focus on the moral implications, fairness, privacy, and societal impact of the study or system, while limitation addresses methodological or contextual constraints.
    (NLP) E.g., "Bias in training data could perpetuate stereotypes in language models."
    (HCI) E.g., "Designing persuasive interfaces risks manipulating users without their awareness."
    (Health Informatics) E.g., "The use of AI for diagnosis must ensure transparency and accountability to maintain patient trust."

  - Ethical Justification: Reasons given to support ethical choices in a study or system.
    (NLP) E.g., "Training data was anonymized to protect user privacy while ensuring model utility."
    (HCI) E.g., "The system prioritizes accessibility features to promote inclusivity for users with disabilities."
    (Health Informatics) E.g., "Patient consent was obtained before using data to ensure adherence to ethical guidelines."

  - Stakeholders/Beneficiaries: Groups or individuals affected by or benefiting from the study or system.
    (NLP) E.g., "Developers of multilingual applications benefit from the improved translation model."
    (HCI) E.g., "End-users of e-learning platforms gain from the enhanced interface usability."
    (Health Informatics) E.g., "Patients with chronic conditions benefit from real-time monitoring and predictive alerts."

1.9 Implications

- **Definition**: Broader applicability or significance and future research suggestion.

- **Subarguments**:
    - Implication: The broader significance or potential impact of the findings
      (NLP) E.g., "Improved sentiment analysis can enhance content moderation on social media platforms."
      (HCI) E.g., "Findings suggest that adaptive interfaces could significantly improve productivity in workplace applications."
      (Health Informatics) E.g., "Real-time health monitoring may reduce hospital readmission rates through early intervention."

    - Risk/Uncertainty: Potential challenges or unknown factors that could affect outcomes.
      (NLP) E.g., "The model may fail when applied to languages not represented in the training data."
      (HCI) E.g., "There is uncertainty about how users will adapt to the new interface in high-stress environments."
      (Health Informatics) E.g., "Reliance on wearable devices introduces the risk of inaccurate health data due to hardware malfunctions."

    - Future Work/Questions: Suggested directions for further research or development.
      (NLP) E.g., "Future work could explore adapting the model for low-resource languages."
      (HCI) E.g., "Further studies are needed to evaluate interface performance in real-world settings."
      (Health Informatics) E.g., "Future research could investigate the integration of genomic data into predictive health models."

1.10 Contradictions to Knowledge

- **Definition**: Points challenging existing knowledge.
- **Subarguments**:
    - Contradiction: Findings that conflict with or challenge prior research.
      (NLP) E.g., "Unlike prior studies, our results show that pretraining on domain-specific data is not always beneficial."
      (HCI) E.g., "Contrary to previous research, we found that users preferred static menus over adaptive ones."
      (Health Informatics) E.g., "Our findings contradict earlier claims that wearable devices always improve patient adherence."

    - Alternative/Counterargument: Differing or opposing perspectives on an issue.
      (NLP) E.g., "An alternative view suggests that rule-based systems are more interpretable than neural models for certain tasks."
      (HCI) E.g., "Some argue that traditional desktop interfaces remain more efficient than mobile-first designs for complex workflows."
      (Health Informatics) E.g., "A counterargument is that patient-reported data might

be less reliable than clinical observations."

- Rebuttal: Responses addressing critiques or opposing views.
  (NLP) E.g., "While rule-based systems are interpretable, they lack the flexibility and scalability of neural models."
  (HCI) E.g., "Although desktop interfaces may be efficient, mobile-first designs better meet the needs of on-the-go users."
  (Health Informatics) E.g., "Despite concerns about reliability, patient-reported data provides valuable insights when combined with clinical records."

2. Annotation Guidelines

11. First read the paragraph and summarize a **Title (event name or event mention)** that best describes the paragraph.

12. With the help from **Title**, find an “Auxiliary-Compound Verb” (“am, is, are” and verbs) from the paragraph that best represents the **Title**. This is called the **“Main Action”** and it has to be an exact quote in the paragraph.

13. Then in this paragraph, find a “main sentence” using the **“Main Action”**, and in most of the cases, “main sentence” is the full sentence that contains the **“Main Action”**. All arguments extracted should be describing this “main sentence” in different ways. Note that there could be other same **“Main Action”** words (as in the English aspect) used in different sentences, you should treat them differently and they should not be the **“Main Action”**. E.g.,

14. There are two additional arguments to find first. The doer of the main action called **“Agent”** and receiver of the action called **“Object”.**

15. Use the **“Main action”** to extract other **arguments**. Then use subarguments above to fit the extracted arguments, and find its argument argument. Don’t infer any information, use only the exact same quote from the given paragraphs. You don’t need to fill all 10 categories for each paragraph, but you are encouraged to cover as much information as possible.

16. You should always try to extract compound nouns only, unless one entire sentence is describing one argument, and the sentence is inseparable in meaning.

17. Following from 6, argument boundaries should not include pronouns unless essential. E.g. "To explore the impact on creators, we present an interview based study of journalists and political commentators on YouTube." As instructed by 6 and 7, choose only an interview based study.

18. In the end, you should annotate **Title, Main action, 10+2 Arguments** (if explicit)

19. Diverse annotations are encouraged! If you check the reconstructed text and find multiple ways of annotating one paragraph, please take note of the paper code, text, and what you think might also be appropriate annotations!

20. Different versions will be kept. **(all are ground truth?) (later use (majority vote) high agreement data 80% to be no dispute.)**

3. Annotation Example

Below is an example that follows the above guideline.

We propose a genetic algorithm (GA) based method for modifying n-best lists produced by a machine translation (MT) system. Our method offers an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics

8. The title of this is “GA based method proposal”
9. The main sentence is: We propose a genetic algorithm (GA) based method for modifying n-best lists produced by a machine translation (MT) system.
10. The **main action** is: “propose”, This can be found in the main sentence
11. Find “**agent**” to be “We”, “**object**” to be “ a genetic algorithm (GA) based method”
12. “for modifying n-best lists produced by a machine translation (MT) system” can be seen as “Goal/Aim” in the “**Purpose**” argument.

    “an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics” is the “Novelty” in the “**result**” argument.

13. Why “for modifying n-best lists produced by a machine translation (MT) system” shouldn’t be separated?

    This is because, the “produced by a machine translation (MT) system” is a strong description of “n-best lists”, and splitting them is wrong, because “produced by a machine translation (MT) system” is not describing the “main action” (which is connected to “GA based method”).

14. A good way to check is by reconstruction and check if the meaning is consistent.

4. Reconstruction Template (self-validate) (pre-validate quality)

(we extract and reconstruct our ground truth as list, and compare with the list from annotators)

Agent + Main Action + Object + In the **context** of _____ + to \ for \ because ____ (**purpose**) + using ____(method) and _____ **evaluation** metric, and the **result** is ___ + and all of the above **implies** ____. This is because ___(**analysis**). This **contradicts** ___. This has some known **limitations**, such as ___. Additionally, ethical issues like _____ are also a part of consideration.

Note that the above template is describing “Main Actions”; All above arguments should be expanded from the “Main sentence”.

Example reconstructed:

We + propose + a genetic algorithm (GA) based method + for “modifying n-best lists produced by a machine translation (MT) system”, and the result is “an innovative approach to improving MT quality and identifying weaknesses in evaluation metrics”.

We compare the meaning of the reconstructed example and found it to be consistent with the original full paragraph.

Todo: Share definition on event trigger.

Improve codebook,

Ambiguous examples, hard cases.

ai2 research AI2 Semantic Scholar Team.

The goal is to understand how people and models think about scientific events.

Then admitting the diverse consideration, design evaluation, and method to guide the models. (even models has different ‘personality’, different models have different way of thinking)

Some known issues:
What if the background and condition appears in the same paragraph? Should we have multiple **Context**?

The boundary of **Object** needs more explanation.

Types of Main actions might need more examples: ACL_23_P_310 Method: “upscale and aggregate” as one action word.

Indicator words

Rebuttal, Novelty, differentiation

Note to self:

Table to count occurrence of the sub-arguments, and merge if appeared rarely

(make students know that their work will be reviewed and given feedback)

(make students understand the benefit and the bigger picture of the research)

Compared to current SoTA codebooks:

1. SciERC (or SciIE)(2018) focus on simple relationships, Ours can handle more in-depth relationships, as well as capture more information:
   E.g., **Explanations** for the main action.
   If using https://nlp.cs.washington.edu/sciIE/annotation_guideline.pdf, they cannot find explanations of scientific entities explained by compound nouns.
   Our ability to annotate longer text span enables the above.
   This codebook focuses more on entity level within one sentence. Our codebook allows cross sentence extraction of information.

2. ACE (2005) codebook is in-depth in defining entities, values, relations, but it is too narrow on the domain of knowledge, and can hardly be applied to our domains. Further, the ACE codebook does not have some important arguments,

relations, and entities that are often found in the scientific domain.

3. SciREX (2020) https://github.com/allenai/SciREX/blob/master/Annotation%20Guidelines.pdf, uses only four entity mentions (material, metric, task, method) to do: entity recognition, entity coreference, and salient entity identification (the more referred entity is a salient one, still coreference task)
   The division of entities in the above way causes great confusion, “DICE score is both a metric and method, but here it behaves like a method because previous annotation make DICE coefficient as metric……”

4. The ACL RD-TEC Annotation Guideline (2016) focus ..