

Semantic Milestones and the SMF-MAP Framework for Semantic Analysis and Verification

Dmytro Lande^[0000-0003-3945-1178],

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, 37, Prospect
Beresteyskyi, Kyiv, Ukraine

dwlande@gmail.com

Abstract

This paper proposes the concept of semantic milestones – short, self-contained, generalizing statements intended for knowledge representation, situation analysis, and decision explanation. The architecture of the SMF-MAP (Semantic Milestone Framework – Multi-Agent Pipeline) multi-agent framework is presented, in which statement generation is separated from evidence verification, the search for alternatives, relationship classification, and editorial control. A taxonomy of nine semantic relationships is developed, taking into account context, quantifiers, and the direction of logical entailment. The results of a sequential single-session model audit are presented, comprising a primary corpus of 90 author-constructed pairs, an additional corpus of 45 pairs, and 90 controlled modifications. For the primary corpus, 87 out of 90 classifications matched the preliminary author-assigned labels before editorial revision; for the additional corpus, 45 out of 45 classifications matched, although substantive reservations remained. In the final modification series, the primary label was preserved for 44 out of 45 paraphrases (97.8%), while exact agreement with the expected label was achieved for 36 out of 45 meaning-changing modifications (80.0%). Ambiguities in reference statements, quantification errors, and the need to specify the direction of logical entailment were identified. All numerical results characterize an internal model audit conducted with access to the authors' expected labels, rather than the independently measured accuracy of SMF-MAP. Full-scale multi-agent runs and blinded expert evaluation remain subjects for future research. In addition, three end-to-end demonstration scenarios are presented to illustrate the generation, critique, and verification of semantic milestones in cybersecurity monitoring, multi-agent coordination, and OSINT tasks; these scenarios do not constitute independent experimental runs.

Keywords: semantic milestones; SMF-MAP; multi-agent systems; large language models; logical entailment; semantic contradiction; paraphrasing; classification robustness; benchmark corpus.

1. Introduction

Reducing large volumes of information to short, meaningful statements is one approach to representing knowledge in a form suitable for human understanding, further analysis, and use in decision-making. However, text compression alone does not guarantee the preservation of meaning, reliability, or practical value. A short statement may be logically correct but trivial; meaningful and appealing but insufficiently substantiated; or convincing under certain

circumstances but false outside them. Moreover, the process of information generalization often results in the loss of the conditions under which a formulated conclusion remains valid.

Particular attention should be paid to the relationships between individual short statements. Two statements may employ almost identical concepts while contradicting each other. At the same time, statements that appear to express opposing positions may prove compatible if they concern different circumstances, time intervals, or aspects of the phenomenon under investigation. For example, the statements “Automation accelerates decision-making” and “Automation complicates decision oversight” do not necessarily contradict each other. The former characterizes the speed of performing procedures, whereas the latter concerns the ability to verify and explain them. Both statements may be valid simultaneously.

This leads to the need to consider short, meaningful statements not as isolated linguistic constructions but as elements of an interconnected knowledge system. For such a system, what matters is not only the formulation of individual conclusions but also the identification of logical, causal, contextual, and functional relationships between them. Of particular importance are contradiction detection, the search for counterexamples, the examination of hidden assumptions, and the determination of the applicability limits of formulated generalizations.

This paper introduces the concept of a semantic milestone—a short, complete, and semantically self-contained statement that captures a significant generalization, observation, prediction, or principle of action. Unlike an ordinary keyword, thematic marker, or summary fragment, a semantic milestone takes the form of an independent proposition. It should not merely designate a particular concept but should express a relationship between concepts, properties, conditions, or consequences.

For example, the phrase “Cybersecurity and monitoring” is a thematic designation rather than a semantic milestone. In contrast, the statement “More alerts do not always mean stronger protection” expresses a complete idea that can be analyzed, refined, and verified. It draws attention to the distinction between the number of recorded events and the actual effectiveness of protective mechanisms. At the same time, its meaning should not be interpreted as denying the benefits of detecting more threats: under certain conditions, additional alerts may indeed indicate improved detection.

The target length of a semantic milestone is up to ten words. This limitation is editorial in nature and is intended to achieve conciseness, expressiveness, and ease of comprehension. It is not a formal logical condition for the correctness of a statement. Priority is given to semantic completeness, clarity, and verifiability. If excessive compression leads to the loss of essential conditions, the relevant qualifications should be preserved in the accompanying semantic representation of the milestone.

Semantic milestones may serve different functions. Some summarize established facts, others describe possible patterns, and still others formulate principles for decision-making. For example, the statement “The number of reports does not imply source independence” reflects a methodological limitation of information analysis. The statement “Local correctness does not guarantee collective consistency” characterizes a potential problem in the interaction of autonomous agents. The statement “Disagreement is sometimes more accurate than consensus” draws attention to the informational value of well-reasoned alternative positions.

Importantly, the value of such statements is determined not only by their conciseness or aphoristic quality. A semantic milestone should result from substantive analysis that considers its origin, logical premises, possible objections, and conditions of validity. Otherwise, the process of generating milestones risks becoming the production of statements that are persuasive in form but insufficiently substantiated.

To address this problem, SMF-MAP (Semantic Milestone Framework – Multi-Agent Pipeline) is proposed as a multi-agent framework for generating, analyzing, critically verifying, and selecting semantic milestones. Its concept is based on the functional division of the process among specialized agents, each responsible for a particular aspect of working with statements. This division makes it possible to separate candidate generation from evaluation, factual verification from logical structure analysis, and final formulation from the search for objections.

A distinguishing feature of the proposed approach is the ability to construct not only a set of individual semantic milestones but also a system of relationships between them. Within such a system, it is possible to identify equivalent and complementary statements, determine the directions of logical entailment, distinguish contradictions from context-dependent differences, and analyze tensions between objectives and resource constraints.

The aim of this paper is to develop the conceptual and formal foundations of SMF-MAP, describe its multi-agent architecture, define an operational classification of semantic relationships, and analyze the results of testing the corresponding rules on specially constructed corpora of statements.

The study encompasses the construction of a corpus of statement pairs, their classification, editorial verification, the development of control examples, and the examination of the robustness of semantic relationships to changes in linguistic formulation. At the same time, it is necessary to distinguish between the conceptually proposed multi-agent architecture and the verification procedures actually performed. The reported results were obtained through single-session model analysis and should not be interpreted as outcomes of independent agent runs or as experimental confirmation of the superiority of SMF-MAP over other approaches.

Thus, the subject of this study is not merely the automatic generation of short statements but the formation of verifiable semantic units that can be used for knowledge representation, the analysis of alternative propositions, and decision support.

2. Theoretical Foundations and Differences from Related Approaches

The formation of semantic milestones is associated with several research areas in natural language processing, knowledge representation, automated reasoning, and multi-agent systems. However, none of these areas individually covers the entire range of tasks involved in creating short, semantically complete, and verifiable generalizations.

The most closely related approach in terms of external characteristics is automatic text summarization. Its purpose is to reduce the length of the original information while preserving its most essential content [1]. Extractive summarization involves selecting fragments of the source

text, whereas abstractive summarization involves generating new linguistic constructions that summarize the information.

However, a semantic milestone is not simply an extremely short summary. A summary generally aims to reflect the content of a document or part of it, whereas a milestone captures a separate, complete statement that can be used independently of the original text. For example, a summary of material on network monitoring may contain information about the number of events, detection methods, and performance indicators. A semantic milestone, in contrast, may formulate the generalization: “Detection sensitivity is not equivalent to threat assessment accuracy.”

Another research area involves semantic vector representations (semantic embeddings), in which words, sentences, or documents are represented in a multidimensional space [2]. Such models make it possible to assess the similarity of text fragments, perform semantic search and clustering, and identify thematically related statements.

However, semantic similarity does not determine the nature of a logical relationship. The sentences “The system detected a threat” and “The system did not detect a threat” may have very similar vector representations, although under identical conditions they describe mutually exclusive states. Conversely, two lexically different expressions may convey equivalent statements.

Therefore, in SMF-MAP, vector representations may be used for the preliminary retrieval of related statements, but they should not replace the examination of their logical and contextual relationships.

A more directly related task is the identification of logical relationships between natural language statements, known as natural language inference (NLI). In a typical NLI setting, three categories are distinguished: entailment, contradiction, and neutrality [3]. This classification makes it possible to determine whether one statement follows from another, whether the statements are incompatible, or whether no sufficiently definite logical relationship can be established between them.

For the analysis of semantic milestones, this set of categories is insufficient. Two statements may neither contradict each other nor stand in a relationship of logical entailment, yet characterize different aspects of the same phenomenon. In other cases, there may be tension between them arising from opposing tendencies, or a trade-off caused by limited resources. Such relationships are independently significant for the analysis of problematic situations.

Consider the statements “Stronger control reduces the risk of abuse” and “Stronger control increases time costs.” Both may be true simultaneously because they describe different consequences of the same process. In the absence of additional conditions, these statements can reasonably be regarded as complementary. However, if it is established that the time available for performing procedures is limited and that stronger control inevitably reduces the time available for core activities, a resource trade-off arises between the corresponding objectives.

Thus, the classification of semantic relationships must take into account not only the content of statements but also the conditions under which they apply. In particular, it is necessary to establish

whether the statements concern the same object, the same time interval, identical constraints, and identical meanings of the concepts used.

Computational argumentation occupies a distinct place in the theoretical foundations of SMF-MAP [4]. Argumentation models make it possible to represent statements, supporting arguments, objections, and relationships between arguments. For the formation of semantic milestones, it is particularly important not only to search for supporting evidence but also to deliberately identify possible counterexamples.

For example, the statement “Automation always increases system reliability” is universal in nature. To refute it, it is sufficient to establish a single valid case in which automation caused an additional error or failure. In contrast, the more cautious formulation “Automation does not eliminate the risk of systemic errors” does not contain an unjustified universal generalization and better reflects the limits of possible conclusions.

Thus, critical verification may serve not only to reject incorrect statements but also to refine their meaning. Identifying a weakness can sometimes transform an excessively categorical statement into a valid semantic milestone.

The multi-agent approach provides an organizational basis for combining the methods described above. Instead of assigning a single language model the simultaneous tasks of generating, substantiating, and ultimately approving a statement, different functions are distributed among specialized agent roles. The Generator proposes candidate statements, the Analyst identifies their semantic structure, the Critic searches for objections, the Verifier checks their validity, and the Editor produces the final linguistic formulation.

At the same time, functional separation does not guarantee the independence of judgments. If all agent roles are implemented using a single language model with a shared context, their responses may be correlated. It is therefore necessary to distinguish between the logical specialization of roles and the actual independence of their execution. Experimental verification of the advantages of the multi-agent approach requires separate runs, controlled configurations, and comparisons with single-stage generation.

The proposed SMF-MAP combines concise knowledge representation, semantic relationship analysis, argumentation-based verification, and the multi-agent organization of information processing. Its distinction from related approaches lies not in the application of a separate new linguistic algorithm, but in an integrated procedure for generating semantic milestones, with explicit representations of context, supporting evidence, objections, and relationships to other statements.

3. Formal Model of a Semantic Milestone

To formalize the process of generating and verifying semantic milestones, it is insufficient to consider them merely as text strings. The same linguistic formulation may acquire different meanings depending on the subject domain, temporal conditions, available evidence, and intended purpose. Therefore, in SMF-MAP, a semantic milestone is represented as a structured object that combines a short statement with the information necessary for its interpretation and verification.

Let a semantic milestone be represented as follows:

$$m_i = (s_i, T_i, K_i, E_i, Q_i, I_i, \tau_i),$$

where s_i is the text of the statement; T_i is the functional type; K_i is the context and scope of applicability; E_i is the evidence and provenance; Q_i is the quality characteristics; I_i is the potential implications for decision-making; and τ_i is the temporal parameters. We distinguish between observational, predictive, and actionable milestones.

The central component is the textual formulation s_i , which must express a complete statement. It is this component that is presented to the user as a short semantic milestone. The remaining components form an accompanying representation that enables semantic analysis and verification. Thus, the conciseness of the external formulation does not require the omission of important qualifications.

The functional type T_i characterizes the purpose of the milestone. Within the proposed model, three main types are distinguished: observational, predictive, and actionable.

Observational milestones summarize established properties, facts, or patterns. For example, “The number of alerts does not determine protection effectiveness” is a statement about the limitations of a particular cybersecurity monitoring indicator. Its correct interpretation requires consideration of the characteristics of the detection system, the number of false positives, and other performance indicators.

Predictive milestones describe the possible consequences of the development of a particular situation. For example, “Accumulating unverified data may increase uncertainty” expresses an assumption about the possible development of an information process. Such a statement requires specifying the conditions under which the corresponding prediction is justified.

Actionable milestones formulate principles for selecting or organizing actions. For example, “Verify source independence before combining evidence” can be used as a practical rule in OSINT analysis.

These types are not necessarily mutually exclusive. A single milestone may contain an observational generalization while also having practical implications. Therefore, the functional type should be regarded as the primary role of a statement in a particular task rather than as an immutable property of its linguistic form.

The context K_i defines the conditions under which a statement can be considered applicable. It may include the subject domain, characteristics of the object, resource constraints, accepted definitions, assumptions, and other circumstances. The absence of explicit context is particularly problematic for short statements, since conciseness is often achieved by omitting conditions.

For example, the milestone “Decision speed does not guarantee decision correctness” is general in nature, but its practical interpretation depends on how speed, correctness, and evaluation criteria

are defined. In cyber incident response tasks, rapid decision-making may be necessary to limit damage, but insufficient verification of information may lead to erroneous actions.

The component E_i contains information about the evidential basis and provenance of the statement. It may include references to sources, computational results, empirical observations, logical arguments, or expert conclusions. It is important to distinguish between statements supported by external data and judgments that are merely conceptually plausible.

The quality characteristics Q_i may include semantic completeness, clarity, consistency, justification, novelty, informativeness, and verifiability. These characteristics should not automatically be combined into a single numerical indicator without defining the scales, aggregation rules, and evaluation procedure. In particular, a high degree of conciseness does not compensate for a factual error, and the expressiveness of a formulation is not evidence of its correctness.

The component I_i characterizes the potential implications of using the milestone in analysis or decision-making. It makes it possible to determine whether the statement may influence the selection of alternatives, refinement of criteria, risk assessment, or identification of directions for further research.

The temporal parameters τ_i specify the time at which the milestone was formulated, the period during which its content remains relevant, or the temporal boundaries of its applicability. This is particularly important for statements related to dynamic processes, including cyber threats, information campaigns, technological changes, and the behavior of adaptive systems.

Individual semantic milestones may be interconnected through different types of relationships. Let

$M = \{m_1, m_2, \dots, m_n\}$ – be the set of generated milestones. Then, their collection, together with semantic relationships and contexts, can be represented as

$$G = (M, R, K),$$

where R – s the set of typed relationships between milestones, and K is the set of contextual characteristics necessary for interpreting the corresponding relationships.

This representation can be regarded as a semantic graph, in which individual milestones constitute the vertices and the relationships established between them constitute the edges. Some of these relationships are directed, particularly logical entailment. Others may be symmetric, such as equivalence or contradiction under identical conditions.

It is fundamentally important that the existence of a semantic relationship does not automatically confirm the truth of the associated statements. For example, two unsubstantiated statements may be logically equivalent, but this does not make them reliable. Similarly, correctly identifying a contradiction between two statements does not determine which of them corresponds to reality.

Consistency of statements is not equivalent to the reliability of knowledge.

Therefore, the formal model provides for the separation of semantic relationships from evidential characteristics. The former describe relationships between meanings, whereas the latter describe the grounds for accepting or rejecting statements.

The use of a semantic graph makes it possible to move from generating isolated aphoristic statements to the systematic analysis of knowledge. Within such a system, it is possible to identify duplicates, detect contradictory generalizations, trace logical consequences, group complementary statements, and identify areas in which the available evidence is insufficient to reach an unambiguous conclusion.

Thus, the proposed formal model combines two levels of semantic milestone representation: a compact textual level intended for human understanding and a structured analytical level designed for verification, comparison, and subsequent use.

4. SMF-MAP Architecture

SMF-MAP (Semantic Milestone Framework – Multi-Agent Pipeline) is designed to transform source information into a set of short, semantically complete, and verifiable semantic milestones. Unlike single-stage statement generation, the proposed architecture involves the sequential execution of several functionally separated procedures, each responsible for a specific aspect of statement formation and evaluation.

The framework is based on a multi-agent pipeline in which the results of the preceding stage are passed to the next stage for analysis, critical verification, refinement, or rejection. In this context, an agent is considered primarily as a functional role with its own task, set of rules, input data, and output data. Such a role may be implemented by a separate language model, a software module, or a specialized invocation of a general-purpose large language model (LLM).

It is important to distinguish between the logical organization of agents and the technical independence of their execution. Distributing functions among several prompts does not, in itself, guarantee the independence of conclusions, particularly when all roles operate within a shared context. Therefore, the architecture provides for the possibility of isolating agent contexts, preserving intermediate results, and logging decisions, although the implementation of these capabilities depends on the specific software configuration.

The overall SMF-MAP architecture, the sequence of input information processing, and the interaction of the framework's functional components are presented in Fig. 1.

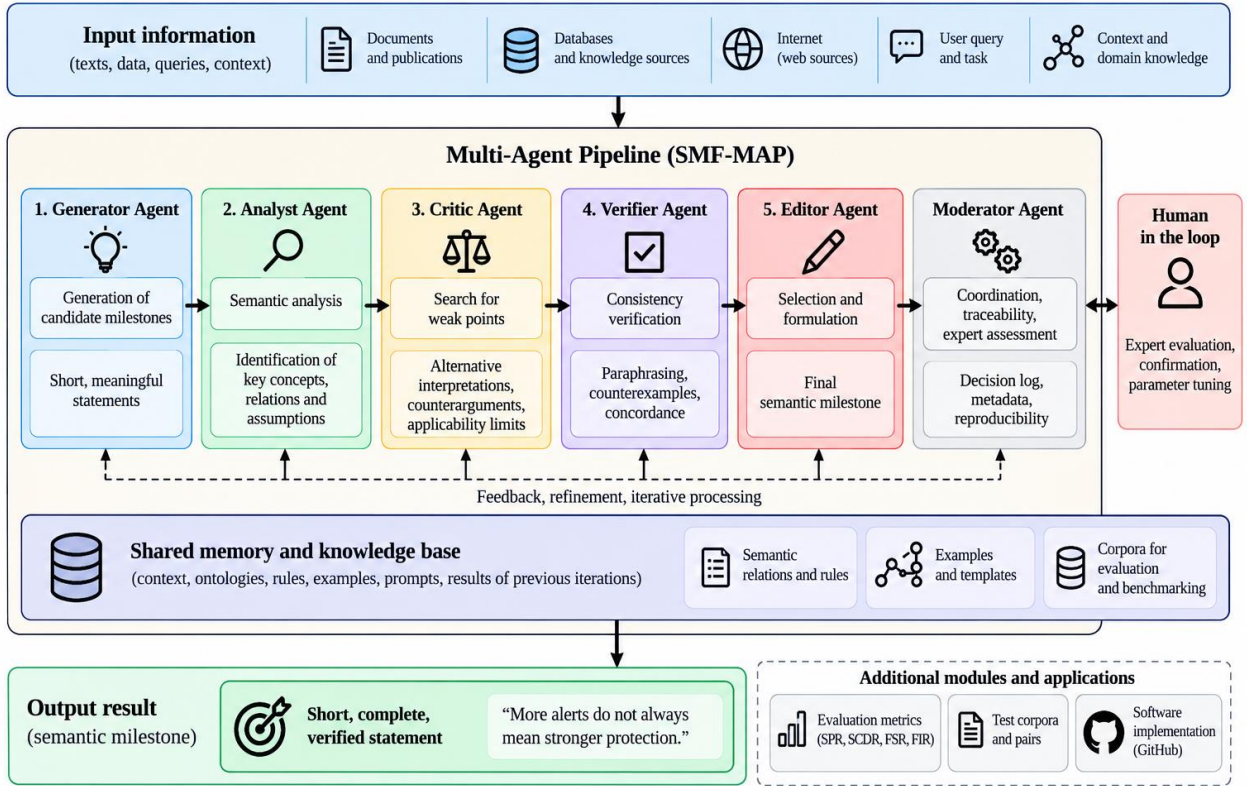


Fig. 1. Overall Architecture of the SMF-MAP Multi-Agent Framework

4.1. Structure and Functional Components of the Framework

As shown in Fig. 1, the SMF-MAP architecture comprises five main functional layers: the formation of the input information environment, multi-agent processing, a shared information base, human interaction, and output generation. Additional tools for evaluation, testing, and reproducibility of procedures are also provided.

The upper part of the figure presents the input information, which may originate from documents and publications, databases, Internet sources, user queries, and domain-specific information resources. This combination of sources makes it possible to use SMF-MAP both for analyzing an individual text and for processing a broader information context.

The user query defines the purpose of generating semantic milestones, the subject domain, the expected nature of the results, and possible constraints. For example, a user may request the formulation of generalizations concerning the effectiveness of cybersecurity monitoring, the identification of contradictions between expert assessments, or the formulation of principles for coordinating autonomous agents.

The input information does not necessarily contain ready-made statements suitable for direct compression. In many cases, it includes facts, event descriptions, statistical indicators,

assumptions, and mutually contradictory assessments. The task of the framework is to establish semantic relationships between these elements and formulate generalizations that do not go beyond the available evidence.

The central part of the figure is occupied by the Multi-Agent Pipeline, which integrates the Generator, Analyst, Critic, Verifier, Editor, and Moderator. Directed information links between the agents indicate the main sequence of candidate milestone processing. The Human in the Loop block is positioned alongside the pipeline, reflecting the possibility of involving a human in the evaluation and approval of results.

Generator Agent

The first functional component is the Generator Agent, which produces a set of candidate semantic milestones based on the input information. Its task is not to immediately select a single best statement, but to generate several semantically distinct alternatives that reflect different aspects of the problem situation.

For example, based on a report of an increase in the number of alerts following a change in detection rules, the Generator may propose the following statements:

“More alerts mean better protection.”

“An increase in alerts may reflect higher detection sensitivity.”

“The number of alerts does not determine protection effectiveness.”

“More alerts do not always mean stronger protection.”

These statements differ in their degree of categorical certainty and semantic implications. The first contains an unjustified causal conclusion, the second offers a possible explanation, the third emphasizes the limitations of an individual indicator, and the fourth expresses a cautious generalization.

The Generator should not independently declare all generated alternatives to be reliable. Its function is to establish a space of candidate statements that will subsequently undergo verification.

Analyst Agent

The second component is the Analyst Agent, which examines the semantic structure of candidate statements. It identifies the main concepts, objects, properties, relationships, assumptions, and conditions of applicability.

For the example above, these concepts include the number of alerts, protection effectiveness, sensitivity of detection rules, actual threats, and false positives. The Analyst must establish that the number of recorded alerts and the level of system protection characterize different aspects of its operation.

Particular importance is attached to identifying hidden transitions from correlation to causation. If the number of alerts increases after detection rules are modified, this does not, in itself, prove either an increase in the number of attacks or an improvement in protection effectiveness. Such a conclusion requires additional data on the nature of the events, the correctness of their classification, and the results of response actions.

The Analyst also establishes relationships between candidate milestones, including equivalence, contradiction, logical entailment, context dependence, and complementarity. In this way, individual statements are gradually transformed into an interconnected semantic configuration.

Critic Agent

The third component, the Critic Agent, systematically searches for weaknesses in candidate statements. Unlike the Analyst, whose primary task is to identify structure and relationships, the Critic attempts to determine the conditions under which a proposed conclusion may be false, overly generalized, or ambiguous.

The main instruments of critical analysis include counterexamples, alternative explanations, the identification of hidden assumptions, and the determination of applicability limits.

For the statement “More alerts mean better protection,” the Critic may formulate the following objection: an increase in the number of alerts sometimes results from excessively sensitive rules that generate a large number of false positives. In such cases, additional alerts increase the workload of operators but do not necessarily improve the actual level of protection.

At the same time, the Critic must also consider an alternative situation in which an increase in the number of alerts is indeed the result of improved attack detection. Thus, the Critic's task is not to mechanically reject every statement, but to identify the conditions under which it can be accepted or rejected.

A counterexample limits a generalization but does not necessarily invalidate its meaning.

Verifier Agent

The fourth component, the Verifier Agent, examines the logical consistency, factual basis, and robustness of candidate milestones to changes in wording.

At this stage, it is important to distinguish between several types of verification. Logical verification determines whether a statement contains internal contradictions or unjustified transitions between propositions. Semantic verification establishes whether meaning is preserved under equivalent paraphrasing. Factual verification involves comparing the statement with external data, when such data are available.

For example, the statement “More alerts do not always mean stronger protection” can be reformulated as “An increase in the number of alerts does not, in itself, prove an improvement in security.” Under an appropriate interpretation, both statements express a similar generalization, although their linguistic forms differ substantially.

However, even successful verification of semantic equivalence does not constitute confirmation of factual truth. The Verifier must separately record which characteristics have been established through logical analysis and which have been confirmed by external sources.

Editor Agent

The fifth component, the Editor Agent, is responsible for the final linguistic formulation of a semantic milestone. Its task is to eliminate redundant words, improve clarity, ensure semantic completeness, and preserve necessary qualifications.

The Editor must avoid situations in which the pursuit of an aphoristic formulation leads to a loss of accuracy. For example, the statement “Alerts do not mean protection” is short but overly categorical, as it ignores the possible relationship between effective event detection and the performance of protective procedures.

The formulation “More alerts do not always mean stronger protection” is more cautious: it rejects the universality of a direct relationship without excluding its existence under particular conditions.

Thus, the Editor does not merely shorten the text but performs semantically controlled generalization in which conciseness is subordinate to accuracy.

Moderator Agent

The final component of the multi-agent pipeline is the Moderator Agent, which coordinates the execution of procedures, controls the sequence of transitions between stages, and ensures that intermediate and final decisions are recorded.

The Moderator may initiate regeneration, additional verification, or the return of a statement to a previous stage if contradictions, insufficient evidence, or ambiguity in wording are identified.

An important function of the Moderator is to ensure traceability. For each final milestone, it is advisable to preserve information about the original query, sources, initial alternatives, critical objections, verification results, and reasons for the editorial choice.

Such an organization makes it possible to reconstruct the reasoning behind the formation of the result and distinguish verified statements from hypotheses or editorial generalizations.

4.2. Shared Memory, Feedback, and Human in the Loop

The lower part of the central block in Fig. 1 presents the Shared Memory and Knowledge Base. This component facilitates information exchange between agent roles and the accumulation of results from previous stages.

The shared information base may include the task context, semantic classification rules, examples of milestones, prompt templates, source information, verification results, and the history of previous iterations.

Particular importance is attached to preserving the rules for establishing semantic relationships. For example, identifying logical entailment requires determining its direction; identifying contradiction requires establishing that the context is the same; identifying a resource trade-off requires the presence of a limited resource; and identifying a counterexample requires a specific case that refutes a general statement.

Shared memory should not be equated with automatically providing every agent with the entire preceding context. Excessive information sharing may lead to dependence between assessments, particularly when the Critic or Verifier effectively repeats the Generator's previous conclusion. Therefore, it is advisable to distinguish between general rules and verified sources, which are accessible to all components, and intermediate judgments, access to which is determined by the experimental protocol.

The dashed arrows in the figure indicate feedback (Feedback, Refinement, Iterative Processing), which enables the repeated processing of candidate milestones. If the Critic identifies a counterexample, the corresponding statement may be returned to the Generator to produce an alternative. If the Verifier identifies ambiguity, the Analyst may clarify the context or definitions of the concepts used. If the Editor cannot shorten a statement without losing its meaning, the Moderator may decide to revise the wording or preserve additional conditions in the accompanying representation.

Thus, the process of generating a semantic milestone is not strictly linear. The sequence of agent roles defines the main processing path, but the actual workflow allows a return to previous stages.

To the right of the Multi-Agent Pipeline in Fig. 1 is the Human in the Loop block, which represents human involvement in evaluating the results. A human may define selection criteria, clarify the subject-domain context, review disputed conclusions, and approve final formulations.

Human involvement is particularly important for milestones containing normative judgments, risk assessments, or recommendations for practical actions. In such cases, even a logically consistent statement may prove unacceptable because of insufficient evidence, domain-specific considerations, or unaccounted-for consequences.

Human involvement does not mean that every statement necessarily requires manual approval. The level of expert participation may be determined by the risk associated with using the result, the degree of uncertainty, and the nature of the task. Nevertheless, the architecture must provide for the possibility of such intervention and the recording of the corresponding decisions.

4.3. Output Generation and Supporting Modules

The lower level of Fig. 1 represents the output of SMF-MAP: a short, complete semantic milestone that has been verified within the scope of the available procedures (Output Result).

The figure provides the following example: “More alerts do not always mean stronger protection.”

This statement demonstrates the main idea of the framework: the final result should be concise without losing an essential semantic qualification. In this case, the phrase “do not always” is

fundamentally important because it prevents a cautious generalization from being transformed into a categorical denial of the relationship between event detection and protection effectiveness.

Along with the milestone text, the system may generate an accompanying structured record containing its functional type, context, sources, verification results, alternative formulations, and semantic relationships with other milestones. Such a record enables the subsequent use of the result in information and analytical systems.

To the right of the output block in Fig. 1 are Additional Modules and Applications. These are not mandatory stages in the generation of every milestone, but they support research into and improvement of the framework.

The first module is intended to evaluate results using specifically defined metrics. In particular, the study of the robustness of semantic relationships uses the Semantic Preservation Rate (SPR), Semantic Change Detection Rate (SCDR), False Stability Rate (FSR), and False Instability Rate (FIR). These metrics characterize the preservation or change of classification decisions under controlled transformations of statements.

The second module contains test corpora and pairs of statements used to verify semantic classification rules. These include examples of logical entailment, contradiction, equivalence, counterexamples, context dependence, and other relationships.

The third module concerns software implementation and reproducibility. It provides for the preservation of prompts, agent configurations, processing rules, and other materials required to repeat the procedures. The author's SMF-MAP repository is available at SMF-MAP GitHub repository (<https://github.com/DmytroLande/SemanticMileStone/>).

The availability of a software implementation and open materials creates the prerequisites for further experimental verification but does not, in itself, confirm the effectiveness of the framework. Such confirmation requires independent series of experiments using fixed corpora, control configurations, and predefined evaluation criteria.

Overall, the architecture presented in Fig. 1 combines the generation of short statements, semantic analysis, reasoned criticism, verification, editorial generalization, and result validation. Its purpose is to move from the one-time generation of aphoristic statements to an organized procedure for producing semantic units with explicit justifications, limitations, and the possibility of subsequent verification.

4.4. Demonstrative Examples of Semantic Milestones

To illustrate the functional purpose of SMF-MAP, twelve original demonstrative examples are presented below. They were formulated during the preparation of this version of the article and are not results of the 360 planned runs, independent expert evaluation, or a comparative experiment. The examples illustrate the transition from an extended statement to a short, self-contained proposition. Each milestone contains no more than ten words.

Example 1. Knowledge and Uncertainty

Original statement: Accumulating facts without verifying their sources may increase confidence but not the justification of a conclusion.

Semantic milestone: “Confidence grows faster than reliability.”

Interpretation: Distinguishing psychological confidence from evidential justification.

Example 2. Knowledge and Uncertainty

Original statement: The absence of contradictions in a report does not guarantee that the report is complete.

Semantic milestone: “Consistency does not necessarily imply completeness.”

Interpretation: Distinguishing internal consistency from completeness of description.

Example 3. Knowledge and Uncertainty

Original statement: New facts may require reconsideration of a previously justified conclusion.

Semantic milestone: “Revising a conclusion does not always imply a previous error.”

Interpretation: Clarifying the nature of scientific revision.

Example 4. Cybersecurity

Original statement: An increase in the number of alerts does not always improve the ability to respond to incidents.

Semantic milestone: “More alerts do not always mean stronger protection.”

Interpretation: Distinguishing alert intensity from response effectiveness.

Example 5. Cybersecurity

Original statement: A system that conceals its own errors makes them more difficult to detect and correct.

Semantic milestone: “An invisible error is more dangerous than a detected one.”

Interpretation: The role of observability in risk management.

Example 6. Cybersecurity

Original statement: Excessive complexity in protective rules sometimes introduces new operational errors.

Semantic milestone: “Protection can create its own vulnerabilities.”

Interpretation: The possibility of secondary risks arising from protection mechanisms.

Example 7. Multi-Agent Systems

Original statement: Several agents may correctly perform their individual local tasks but collectively produce an undesirable result.

Semantic milestone: “Local correctness does not guarantee collective consistency.”

Interpretation: Distinguishing agent correctness from a system-level property.

Example 8. Multi-Agent Systems

Original statement: Disagreement between agents may reveal a hidden error in task formulation.

Semantic milestone: “Disagreement is sometimes more accurate than unanimity.”

Interpretation: The productive role of reasoned alternatives.

Example 9. Multi-Agent Systems

Original statement: A large number of agents using the same source does not provide independent confirmation.

Semantic milestone: “The number of voices does not equal evidence independence.”

Interpretation: Distinguishing the number of responses from the independence of their origins.

Example 10. Decision-Making

Original statement: The choice of an optimal action depends on objectives and constraints, which may change.

Semantic milestone: “The best decision depends on the question asked.”

Interpretation: The context-dependent nature of optimality.

Example 11. Decision-Making

Original statement: A decision made too quickly may save computation time but increase the cost of error.

Semantic milestone: “Decision speed does not determine its value.”

Interpretation: Distinguishing decision speed from decision quality.

Example 12. Decision-Making

Original statement: When two criteria are incompatible under limited resources, fully satisfying one may reduce the degree to which the other is satisfied.

Semantic milestone: “Trade-offs arise where goals share a resource.”

Interpretation: The conditions under which resource trade-offs emerge.

4.5. Three End-to-End Scenarios for Generating and Verifying Semantic Milestones

The scenarios presented below are transparent methodological reconstructions of the sequence of logical roles in SMF-MAP v1.1.

The sequence of semantic milestone generation, semantic analysis, critical verification, and final selection is illustrated in Fig. 2 using a cybersecurity monitoring task as an example. The diagram represents the logic of interaction between agent roles and is intended for demonstration purposes.

These are not logs of actual independent agent runs, nor do they constitute empirical verification of effectiveness. For each scenario, the input, initial candidates, alternative judgment, critical objection, verification status, and editorial decision are presented. According to the framework specification, agent roles may be executed sequentially by a single LLM; such a reconstruction does not demonstrate their independence.

Scenario 1. Cybersecurity Monitoring: Number of Alerts and Detection Quality

SITUATION mode. Input situation: Following an increase in the sensitivity of an intrusion detection system, the number of alerts increased. No information is provided on the proportion of true positives, missed attacks, or response time. The task is to formulate a meaningful milestone without equating the number of alerts with protection effectiveness.

The Coordinator specifies the observational type O and the scope of the conclusion. The Context Agent distinguishes the established fact of an increase in the number of alerts from unknown accuracy metrics. The Generator proposes three candidates: “More alerts mean better protection”; “The number of alerts is not equivalent to protection quality”; and “Detector sensitivity changes the analyst's workload.” The first candidate is rejected because of an unjustified causal generalization; the third requires qualifications concerning the nature of the alerts and the workflow.

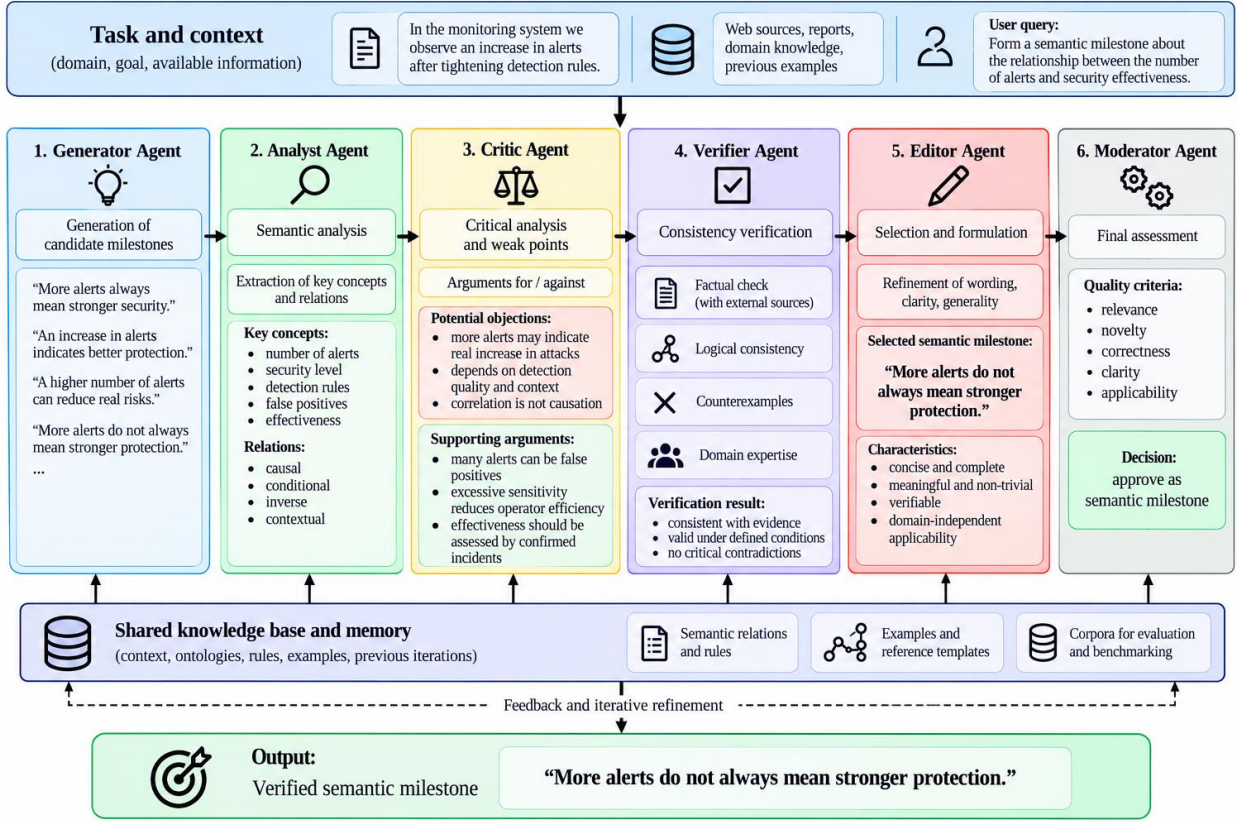


Fig. 2. Generation and Verification of a Semantic Milestone Using SMF-MAP

The Alternative Judgments Agent formulates the statement: "Additional alerts may reveal previously missed attacks." The Semantic Analyst identifies the relationship as `CONTEXT_DEPENDENT` with respect to the statement concerning the limited informativeness of alert counts: if genuinely new detections are present, an increase in their number may be beneficial, but the increase alone does not establish this. The Critic proposes a counter-scenario: a sharp increase in false positives raises the workload without improving protection. The Verifier assigns the status `conditional`, since the input situation contains no independent measurements.

The Editor retains milestone O1: "More alerts do not always mean stronger protection." Decision: `retain_with_qualification`. Practical verification would require precision, recall, false positive rate, and response time measurements before and after the configuration change. The milestone is a conditional generalization, not a conclusion about the actual effectiveness of a particular IDS.

Scenario 2. Multi-Agent Coordination: Local Correctness and Collective Outcomes

SITUATION mode. Input situation: Two agents perform local tasks according to predefined rules; the first minimizes processing latency, while the second maximizes verification completeness. The joint system has a limited time resource. No numerical data or evidence of failure is available. The task is to describe a possible conflict of objectives without claiming that it has already occurred.

The Context Agent identifies different objective functions and a shared constraint. The Generator proposes: "Correct agents always form a correct system"; "Local correctness does not guarantee

collective consistency”; and “Shared constraints may bring autonomous objectives into conflict.” The first candidate is rejected: correctness with respect to a local objective does not imply satisfaction of the global objective. The second is a generalizing observational statement of type O; the third is a conditional predictive statement of type P.

The Alternative Judgments Agent proposes: “Coordinated constraints allow agents to operate without conflict.” Its relationship to the third candidate is `CONTEXT_DEPENDENT` rather than `CONTRADICTION`, since the coordination conditions differ. The Critic notes that divergence of objectives does not necessarily produce an actual conflict; such a conflict requires a functionally significant incompatibility of actions or outcomes. The Verifier marks both generalizations as `conditional`; no actual conflict has been measured.

The Editor retains O2: “Local correctness does not guarantee collective consistency.” An additional predictive milestone, P2, is formulated: “Shared constraints may bring autonomous objectives into conflict.” Decisions: `retain`; `retain_with_qualification`. Verifiable indicators for a future experiment include the frequency of global constraint violations, the number of replanning operations, and the proportion of successfully coordinated actions.

Scenario 3. OSINT: Multiplicity of Reports and Source Independence

TEXT/SITUATION mode. Input training passage: “Five channels disseminated the same report about an incident. All of them refer to a single original publication. No independent documentary confirmation is available.” This is a deliberately constructed example, not a description of a real incident. The task is to determine what can legitimately be stated about the evidential strength of the reports.

The Context Agent identifies a single original source and five dissemination channels. The Generator proposes: “Five reports independently confirm the incident”; “Repeating a source does not create independent confirmation”; and “The number of publications may conceal a common origin.” The first candidate contradicts the specified condition of a shared source and is rejected. The second is directly supported by the input text; the third is a broader conditional generalization.

The Alternative Judgments Agent formulates: “Five channels may increase the accessibility of the report.” The relationship is `COMPLEMENT`: dissemination and independent evidential strength are different characteristics. The Critic examines the risk of an excessive conclusion: the absence of independent confirmation does not prove that the report is false. The Verifier assigns the status `supported` to the conclusion about common origin strictly within the training passage and `conditional` to the general rule.

The Editor retains O3: “The number of voices does not equal evidence independence.” The actionable milestone A3 is: “Verify provenance before counting confirmations.” Decisions: `retain`; `retain`. The next practical step is to construct a provenance graph of the reports, identify the original sources, and distinguish copies from independent evidence. This scenario demonstrates the transition from an aphoristic conclusion to a verifiable analytical action.

Comparison of Scenarios and Limitations of the Demonstration

In the first scenario, critical verification narrows an unjustified causal inference; in the second, it distinguishes potential incompatibility of objectives from an actually observed conflict; in the third, it separates information dissemination from independent confirmation. In all three cases, the milestones are concise, while the necessary qualifications are preserved in the accompanying metadata. Accordingly, the conciseness of the output statement does not eliminate the need for context, evidence, and counterexamples.

The scenarios do not provide a sample-based assessment of generation quality, establish inter-agent agreement, or demonstrate superiority over a single-stage LLM baseline. Such verification requires fixed input data, separate logs of agent calls, independent evaluation, and the C0–C3 comparison planned in the research protocol.

4.6. Illustration of the Verification of Alternative Interpretations

The verification of semantic milestones requires more than establishing the grammatical correctness or apparent plausibility of a short statement. A milestone may be meaningful and logically consistent under one interpretation while becoming misleading or unjustified under another. This problem is particularly significant for concise formulations, since brevity is often achieved by omitting contextual conditions, implicit assumptions, and qualifications.

Consequently, the verification procedure should examine not only the explicit content of a milestone but also the range of interpretations that may reasonably be associated with it.

Consider the milestone “Disagreement is sometimes more accurate than unanimity.” At first glance, this statement may appear to suggest that disagreement is generally preferable to consensus. Such an interpretation would be incorrect. The statement does not establish a universal advantage of disagreement; rather, it identifies the possibility that an independently justified alternative position may be more reliable than a consensus formed under shared assumptions or common sources of error.

For example, several analytical agents may reach the same conclusion because they rely on an identical information source or reproduce the same reasoning pattern. Their agreement does not necessarily provide independent confirmation. Another agent, using different evidence or identifying an overlooked assumption, may formulate a dissenting conclusion that better corresponds to the available facts. Under these conditions, disagreement can have greater informational value than unanimity. However, the opposite situation is also possible: a dissenting agent may rely on incomplete information or incorrect reasoning, while the consensus may be supported by independent and mutually consistent evidence.

The verification task is therefore not to determine whether disagreement or unanimity is inherently superior, but to establish the conditions under which the milestone remains justified. The Critic Agent should examine counterexamples and alternative explanations, while the Verifier Agent should distinguish logical plausibility from factual support. The resulting interpretation must preserve the qualification expressed by the word “sometimes” and associate the milestone with an explicit context concerning the independence and quality of evidence.

A similar problem arises with the milestone “More alerts do not always mean stronger protection.” Its interpretation depends on distinguishing the number of generated alerts from the actual effectiveness of cybersecurity mechanisms. An increase in alerts may result from improved detection sensitivity, changes in monitoring rules, additional malicious activity, or a higher false-positive rate. These explanations are not mutually exclusive, and the observed increase alone does not establish which of them is correct.

A possible objection is that additional alerts may indeed indicate improved detection of previously unnoticed attacks. This objection does not contradict the milestone because the statement does not deny the possibility of such improvement. Instead, it rejects the universal inference that every increase in alert volume necessarily corresponds to stronger protection. The verification procedure should therefore preserve the distinction between an increase in observable events and an improvement in security outcomes.

This example illustrates the importance of analyzing quantifiers and modal expressions. Replacing “do not always” with “never” would substantially change the meaning of the milestone. The original formulation denies a universal relationship, whereas the modified formulation would deny the relationship in every possible case. Such a transformation cannot be treated as semantically equivalent, even though the two statements contain nearly identical words.

Alternative interpretations may also arise from differences in the meaning of individual concepts. For instance, the milestone “Local correctness does not guarantee collective consistency” requires distinguishing the correctness of individual agent actions from the properties of the overall multi-agent system. An agent may correctly optimize its assigned objective while contributing to an undesirable collective outcome. However, this does not imply that every system of individually correct agents is necessarily inconsistent. The milestone expresses the absence of a logical guarantee rather than the inevitability of collective failure.

These examples demonstrate that semantic verification must account for the distinction between a statement's literal meaning, its possible interpretations, and the evidence supporting its application. In SMF-MAP, this distinction can be represented through the context component, supporting evidence, and semantic relationships associated with each milestone. The short textual formulation remains accessible to the user, while the accompanying representation preserves the qualifications required for correct interpretation.

The verification of alternative interpretations also provides a mechanism for improving milestone formulations. If a statement admits several incompatible interpretations, the Editor Agent may introduce a qualification, replace an ambiguous concept, or select a more precise formulation. If the ambiguity cannot be resolved without additional information, the system should preserve the uncertainty rather than impose an unjustified interpretation.

Thus, the verification of alternative interpretations is not merely an additional quality-control procedure. It is a necessary component of transforming concise statements into reliable elements of knowledge representation. The purpose of this procedure is not to eliminate every possible interpretation but to distinguish justified interpretations from those that exceed the available evidence.

4.7. From Demonstrative Milestones to a Verifiable Experiment

The examples presented in the preceding subsections illustrate the conceptual principles of SMF-MAP and the functional distribution of tasks among its agents. However, the successful construction of meaningful examples does not, by itself, demonstrate the effectiveness of the proposed framework. A distinction must therefore be made between the explanatory value of a demonstration and the empirical evidence required to evaluate a computational method.

A demonstrative milestone is typically selected or constructed to illustrate a particular semantic property. Such examples are useful for explaining the operation of the Generator, Analyst, Critic, Verifier, and Editor agents, as well as for clarifying the relationships between alternative statements. Nevertheless, their selection may introduce methodological bias. Examples that are especially clear, expressive, or easy to verify do not necessarily represent the variety of statements encountered in real information environments.

A reproducible evaluation of SMF-MAP requires the preparation of a predefined set of input tasks and reference materials. These tasks should represent different conditions under which semantic milestones may be generated, including the analysis of source texts, descriptions of problem situations, and general thematic requests. The input materials should vary in complexity, ambiguity, completeness, and the degree of evidential support available.

The evaluation should also distinguish between the quality of individual milestones and the quality of the procedure used to generate them. At the level of individual statements, relevant characteristics include semantic completeness, clarity, conciseness, factual justification, and the preservation of essential contextual conditions. At the procedural level, the evaluation should consider the ability to identify unsupported generalizations, detect contradictions, generate meaningful alternatives, and explain the reasons for accepting or rejecting candidate statements.

An important requirement is the separation of generation and evaluation. If the same language model constructs the examples, assigns reference labels, and evaluates its own outputs within a shared context, the resulting agreement cannot be interpreted as an independent measure of correctness. Information about expected outcomes may influence subsequent decisions, even when the evaluation procedure formally distinguishes different agent roles.

To reduce this limitation, future experiments should employ fixed corpora, independent reference annotation, and evaluation conditions in which the models being tested do not have access to expected labels. Expert assessments should be collected independently, and disagreements between evaluators should be documented and resolved according to predefined rules. The resulting reference data would provide a more reliable basis for estimating classification performance and evaluating the quality of generated milestones.

The multi-agent architecture should also be evaluated against simpler configurations. In particular, a comparison between single-stage LLM generation and configurations incorporating additional analytical, critical, and verification functions would make it possible to determine whether the separation of agent roles produces measurable improvements. Such comparisons should use identical input tasks and clearly defined evaluation criteria.

The experimental protocol should record the model versions, prompts, generation parameters, execution sequence, intermediate outputs, and final decisions. Repeated runs would allow researchers to examine the stability of results and distinguish systematic improvements from variations associated with individual model responses. The contribution of particular agents could subsequently be investigated through controlled ablation experiments.

A separate evaluation task concerns the robustness of semantic relationships under linguistic transformations. A reliable classification procedure should preserve the identified relationship when a statement is paraphrased without changing its meaning. Conversely, it should recognize modifications that alter logically significant elements, including quantifiers, negation, temporal conditions, numerical thresholds, or the direction of logical entailment.

The metrics introduced in this study, including the Semantic Preservation Rate (SPR), Semantic Change Detection Rate (SCDR), False Stability Rate (FSR), and False Instability Rate (FIR), provide an operational basis for examining these properties. However, their interpretation depends on the correctness and independence of the reference labels. Agreement with an incorrectly assigned label does not demonstrate semantic correctness, while disagreement may sometimes indicate that the classification procedure has identified a weakness in the reference example.

The distinction between demonstrative and experimental results is also important when considering the broader intellectual context of semantic milestones. Concise generalizations have a long history in philosophical, literary, and scientific discourse. Comparisons with aphoristic traditions, including the rubaiyat of Omar Khayyam and Laozi’s Tao Te Ching, may help clarify the expressive characteristics of short statements. Nevertheless, such comparisons require explicit selection criteria, attention to translation and cultural context, and independent assessment. They should not be treated as evidence of the superiority of automatically generated milestones without an appropriate experimental design.

The present study therefore regards the examples in Section 4 as methodological demonstrations rather than independent evidence of SMF-MAP performance. The internal model-based audit reported in subsequent sections examines the behavior of semantic classification rules on constructed corpora, but it does not replace a controlled evaluation of the complete multi-agent framework.

The transition from demonstration to experimentation is essential for determining which properties of SMF-MAP follow from its conceptual organization and which can be confirmed empirically. A successful future evaluation should establish not only whether the framework produces concise and meaningful statements, but also whether its verification procedures improve their reliability, whether its classifications remain stable under appropriate transformations, and whether the resulting semantic milestones provide useful explanations for human decision-making.

5. Operational Taxonomy of Semantic Relationships

The generation of individual semantic milestones does not exhaust the capabilities of SMF-MAP. An equally important task is to establish semantic relationships between them, since these relationships make it possible to move from a collection of short statements to a structured representation of knowledge. Two statements may be equivalent, contradict each other, stand in a

relationship of logical entailment, characterize different aspects of the same phenomenon, or express opposing tendencies. At the same time, apparent linguistic similarity does not always correspond to the nature of their semantic relationship.

Within SMF-MAP, an operational taxonomy comprising nine basic classes of semantic relationships is proposed. Its purpose is not merely to describe possible relationships between statements but also to define the rules by which these relationships can be established, verified, and reproduced. Unlike purely descriptive classifications, the operational taxonomy specifies recognition criteria, conditions of applicability, characteristic examples, and rules for distinguishing each class from related classes.

The need for such detail arises from the fact that semantic relationships often share common characteristics. For example, a counterexample may simultaneously indicate a contradiction between a general statement and a specific fact. A resource trade-off may be accompanied by tension between objectives. Two statements that appear contradictory may prove compatible once the time, object, or conditions under which they hold true are clarified. Therefore, identifying the primary class requires analyzing not only the wording of the statements but also the logical structure of their relationship.

Let A and B be two statements between which a semantic relationship must be established. Let K denote the set of contextual conditions defining the subject domain, objects, temporal parameters, accepted definitions, and other circumstances relevant to interpretation. The classification task then consists of determining the relationship

$$r(A, B | K) \in R,$$

$$r(A, B | K) \in \mathcal{R},$$

where R is the set of nine basic classes of semantic relationships.

In cases where the available contextual conditions are insufficient to determine the relationship unambiguously, the procedure must allow abstention from classification. Such an outcome is not regarded as a separate semantic class but indicates insufficient information for making a justified decision.

Table 1 presents the basic taxonomy of semantic relationships used in SMF-MAP.

Table 1. Operational Taxonomy of Semantic Relationships in SMF-MAP

Class	Distinguishing Criterion	Representative Example
CONTRADICTION	Impossibility of both statements being true simultaneously under identical conditions	“Access is permitted” — “Access is denied” for the same request
COUNTEREXAMPLE	A specific case refutes a universal statement	“All backups are usable” — “This backup cannot be restored”
EQUIVALENCE	Bidirectional logical entailment	“No node has failed” — “All nodes remain operational,” under identical definitions
ENTAILMENT	Unidirectional logical entailment with an explicitly specified direction	“All five checks were successful” — “At least one check was successful”
CONTEXT_DEPENDENT	An apparent discrepancy is explained by differences in conditions	“The algorithm is fast on small datasets” — “The algorithm is slow on large datasets”
TRADE_OFF	A necessary choice between alternatives under a limited shared resource	“More memory for caching” — “Less memory for analysis”
TENSION	Opposing tendencies without a demonstrated necessary resource trade-off	“Autonomy accelerates actions” — “Autonomy complicates control”
COMPLEMENT	Compatible statements describing different aspects of the same phenomenon	“Encryption protects content” — “Authentication verifies the participant”
UNRELATED	Absence of a significant established relationship	“The server has two processors” — “The conference will take place in autumn”

The examples provided are illustrative. Their classification assumes a common understanding of the concepts used and the absence of additional conditions that would alter the meaning of the statements.

5.1. Logical Contradiction (CONTRADICTION)

The CONTRADICTION relationship is established between statements that cannot be simultaneously true under identical conditions. This refers to the incompatibility of their meanings with respect to the same object, time interval, and aspect under consideration.

Formally, contradiction can be represented as

$$K \models \neg(A \wedge B).$$

This expression means that, under the accepted contextual conditions K , statements A and B cannot both be true.

For example, consider the following statements:

A: “At 12:00, the server was operational.”

B: “At 12:00, this server was not operational.”

If both statements use the same definition of operational status, they are contradictory.

However, changing the time in the second statement alters the situation:

A: “At 12:00, the server was operational.”

B: “At 12:05, the server was not operational.”

Both statements can now be true because they describe different points in time. The CONTRADICTION relationship no longer follows directly from their content.

Particular attention must be paid to the scope of negation. For example, the statement “Not all agents completed the task” is not equivalent to the statement “No agent completed the task.” The former denies that successful completion was universal, whereas the latter denies the existence of even a single successful case.

Therefore, establishing a contradiction requires verifying the identity of the objects, temporal conditions, definitions, and logical structure of the statements.

Opposing formulations do not, in themselves, establish a contradiction between statements.

5.2. Counterexample (COUNTEREXAMPLE)

The COUNTEREXAMPLE relationship characterizes a situation in which a specific case refutes a general statement. Unlike an ordinary contradiction, the structure of the relationship between a universal statement and a specific instance of evidence is of particular importance.

Let the general statement have the form

$$A: \forall x \in D, P(x),$$

where D is the domain under consideration and $P(x)$ is a particular property.

A counterexample is established by identifying an object x_0 for which

$$x_0 \in D, \neg P(x_0).$$

For example:

A: “All messages generated by the monitoring system are reliable.”

B: “Message No. 125 contains erroneous data.”

If Message No. 125 belongs to the system in question and the error has been confirmed, the second statement refutes the first.

In SMF-MAP, a counterexample is treated as a separate class because it characterizes not only logical incompatibility but also a method for testing a generalization. Its identification is directly relevant to the work of the Critic Agent.

However, not every individual case constitutes a valid counterexample. It is necessary to establish that the object in question belongs to the scope of the general statement, that its properties have been properly determined, and that the evidence presented actually refutes the universal statement.

For example, a report of an error in another system does not refute a statement concerning a particular monitoring system. Similarly, an unverified assumption about the possibility of an error does not constitute a factual counterexample.

The COUNTEREXAMPLE relationship may overlap with CONTRADICTION. In such cases, COUNTEREXAMPLE should preferably be assigned as the primary label if a specific case is used to refute a universal generalization. CONTRADICTION may be indicated as a secondary relationship if appropriately justified.

A single counterexample is sufficient to refute a universal statement.

5.3. Semantic Equivalence (EQUIVALENCE)

The EQUIVALENCE relationship is established between statements that have identical truth conditions within a defined context. Its formal criterion is bidirectional logical entailment:

$$A \Rightarrow B, B \Rightarrow A.$$

Equivalence does not require identical linguistic forms. On the contrary, establishing equivalence is particularly important for recognizing different formulations of the same meaning.

For example:

A: “None of the five nodes completed the check with an error.”

B: “All five nodes completed the check without an error.”

Provided that every check was completed and the same definition of an error is used, these statements may be regarded as equivalent.

At the same time, linguistic similarity does not guarantee equivalence. Consider the following statements:

A: “All five nodes passed the check.”

B: “At least one of the five nodes passed the check.”

The first statement entails the second, but the reverse entailment does not hold. Therefore, the relationship between them is ENTAILMENT rather than EQUIVALENCE.

When establishing equivalence, it is necessary to consider quantifiers, numerical values, modal constructions, and implicit assumptions. For example, “The system can detect a threat” and “The system detects a threat” are not equivalent: the first statement describes a capability, whereas the second may describe an actual or regular action.

The EQUIVALENCE relationship is of practical importance for eliminating duplicate semantic milestones, consolidating related formulations, and verifying the robustness of meaning under linguistic transformations.

Different words can express identical propositions.

5.4. Logical Entailment (ENTAILMENT)

The ENTAILMENT relationship characterizes a situation in which the truth of one statement, under specified conditions, guarantees the truth of another, while the reverse entailment does not hold.

Unlike equivalence, this relationship is directional. The SMF-MAP v1.1 rules specify two subclasses.

The first subclass is ENTAILMENT_A_TO_B:

$$A \Rightarrow B, B \nRightarrow A.$$

The second subclass is ENTAILMENT_B_TO_A:

$$B \Rightarrow A, A \nRightarrow B.$$

These designations do not constitute two additional classes in the taxonomy. They specify the direction of the single basic ENTAILMENT relationship.

For example:

A: “All seven agents completed processing the request.”

B: “At least one agent completed processing the request.”

Provided that the seven specified agents exist, statement A entails B. However, the fact that one agent completed processing does not imply that all seven agents did so.

Therefore, the relationship is classified as ENTAILMENT_A_TO_B.

If the statements are reversed, the direction of entailment changes, although the underlying semantic relationship remains the same.

Directionality is particularly important when constructing semantic graphs. If a logical entailment relationship is established between vertices, the corresponding edge must be directed. Otherwise, the system may incorrectly interpret a weaker statement as a basis for a substantially stronger conclusion.

For example, “All sources have been verified” entails “Some sources have been verified,” provided that the set of sources is nonempty. However, the second statement does not justify the conclusion that all sources have been verified.

Thus, explicitly specifying the direction of ENTAILMENT is a necessary condition for the correct use of semantic relationships in automated reasoning procedures.

Logical entailment has a direction that cannot be ignored.

5.5. Context Dependence (CONTEXT_DEPENDENT)

The CONTEXT_DEPENDENT relationship is established between statements whose apparent discrepancy can be explained by differences in their conditions of applicability.

Unlike CONTRADICTION, this relationship does not involve a necessary incompatibility of truth values. Both statements may be true simultaneously if they refer to different contexts.

For example:

A: “The algorithm provides high processing speed for small datasets.”

B: “The algorithm operates slowly on large datasets.”

These statements do not contradict each other because processing speed depends on data volume.

Context dependence may be determined by task size, available resources, temporal conditions, environmental characteristics, system operating modes, or accepted definitions.

In cybersecurity, particularly illustrative situations arise when a protective mechanism demonstrates high effectiveness under certain conditions but insufficient effectiveness under others. For example, a detection system may achieve high accuracy for known attack types while exhibiting lower effectiveness against previously unknown attack scenarios.

Classifying a relationship as CONTEXT_DEPENDENT requires more than merely assuming that some additional conditions might explain the discrepancy. A specific difference between contexts must be established on the basis of the statements themselves or the data provided.

If no such difference can be identified, the procedure must abstain from assigning an unambiguous classification or request clarification.

A contradiction sometimes disappears when the conditions are clarified.

5.6. Resource Trade-Off (TRADE_OFF)

The TRADE_OFF relationship characterizes a situation in which improving one indicator, achieving one objective, or expanding one capability requires limiting another because of a shared scarce resource.

The defining characteristic of this class is not merely the opposition of consequences but an established mechanism that necessitates a choice.

For example, suppose a system has a fixed amount of random-access memory:

$$M_{cache} + M_{analysis} \leq M_{total},$$

where M_{cache} is the memory allocated for caching, $M_{analysis}$ is the memory allocated for analytical procedures, and M_{total} is the total available memory capacity.

If the total memory capacity is fixed and fully allocated, increasing the resources allocated to caching requires reducing those available for analysis.

The corresponding statements may take the following form:

A: “Increasing cache size accelerates repeated accesses.”

B: “Increasing cache size reduces the memory available for analysis.”

Under the specified conditions, a resource trade-off exists between the corresponding functions.

However, if the system has a substantial reserve of unused memory, increasing the cache size does not necessarily require reducing other resources. In such a case, the TRADE_OFF relationship cannot be established without additional conditions.

A resource trade-off may involve not only memory but also time, energy, bandwidth, computing capacity, budget, or human attention.

In SMF-MAP, the TRADE_OFF relationship must be distinguished from TENSION. If there is a general tension between objectives but a necessary resource trade-off has not been demonstrated, the primary label should be TENSION.

A trade-off arises where choice is constrained by resources.

5.7. Semantic Tension (TENSION)

The TENSION relationship describes a situation in which two statements reflect opposing tendencies, objectives, or consequences, but their interaction does not constitute a demonstrated logical contradiction or a necessary resource trade-off.

For example:

A: “Agent autonomy accelerates task execution.”

B: “Agent autonomy complicates centralized control.”

Both statements may be true simultaneously. However, they indicate opposing consequences of increased autonomy.

Unlike TRADE_OFF, this relationship does not establish that improving speed necessarily requires a corresponding reduction in the level of control. Architectural solutions may exist that allow autonomy to be increased without substantially reducing controllability.

Semantic tension may arise between speed and caution, openness and confidentiality, autonomy and centralization, accuracy and generalization, or stability and adaptability.

Identifying such relationships is particularly useful when analyzing complex systems, as it makes it possible to detect potential problems before an explicit conflict emerges.

At the same time, TENSION should not automatically be interpreted as a negative property. Tension between different requirements may serve as a source of system development, stimulate the search for new solutions, or contribute to refining evaluation criteria.

Tension between objectives does not always imply an inevitable trade-off.

5.8. Semantic Complementarity (COMPLEMENT)

The COMPLEMENT relationship is established between compatible statements that describe different but semantically related aspects of the same object, process, or situation.

For example:

A: “Encryption ensures the confidentiality of transmitted data.”

B: “Authentication enables verification of the identities of communication participants.”

Both statements characterize different functions of information security mechanisms. They are not equivalent, neither entails the other, and they do not contradict each other. At the same time, considering them together provides a more complete description of secure information exchange.

Unlike TENSION, the COMPLEMENT relationship does not involve a pronounced opposition of tendencies. Unlike UNRELATED, it characterizes a meaningful semantic connection between statements.

Semantic complementarity is particularly important when forming collections of milestones intended to provide a comprehensive description of a problem situation. It allows different observations to be combined without artificially reducing them to a single generalization.

For example, when analyzing a cyber incident, one milestone may characterize the technical indicators of an attack, another may describe user behavior, and a third may address organizational constraints on incident response. Together, they form a multidimensional representation of the situation.

However, the compatibility of statements alone is not sufficient to establish COMPLEMENT. A meaningful semantic connection must be identified through which their joint consideration enriches the description of the object.

Completeness of knowledge often emerges from combining different perspectives.

5.9. Absence of an Established Semantic Relationship (UNRELATED)

The UNRELATED class is used for pairs of statements between which no significant semantic relationship has been established within the provided context.

For example:

A: “The server is equipped with two processors.”

B: “The scientific conference will take place in September.”

In the absence of additional context, these statements characterize different objects and do not form a semantically significant relationship.

Importantly, UNRELATED does not imply that establishing any relationship is absolutely impossible. Under additional conditions, even apparently independent statements may prove to be related. For example, if the server mentioned is used to support the conference, a domain-specific connection may arise between the statements.

However, the classifier must not invent such a context independently. A relationship is established only on the basis of the available statements and explicitly specified conditions.

UNRELATED must also be distinguished from NEEDS_CONTEXT. In the former case, the available information is sufficient to conclude that no significant relationship has been established within the specified task. In the latter case, a specific uncertainty prevents the selection of one of the possible semantic classes.

For example, if two statements contain the word “key,” but one refers to a cryptographic key and the other to a physical key, classification depends on whether the meaning of the term has been established. If the meanings are defined and no domain-specific relationship exists, UNRELATED may be assigned. However, if the ambiguity of the term remains unresolved, NEEDS_CONTEXT may be more appropriate.

The absence of an established relationship does not prove its impossibility.

5.10. Priority Rules and Distinguishing Related Classes

The operational taxonomy provides that the primary semantic relationship is determined according to the most specific criterion directly supported by the content of the statements.

In particular, if a specific case refutes a universal statement, priority is given to COUNTEREXAMPLE, even if a logical incompatibility also exists between the statements.

If bidirectional logical entailment is established, EQUIVALENCE is assigned. If entailment is unidirectional, ENTAILMENT is assigned with an explicit indication of its direction.

If statements appear contradictory but their discrepancy can be explained by explicitly established contextual differences, priority is given to CONTEXT_DEPENDENT.

If statements characterize opposing tendencies, it is necessary to examine whether a limited shared resource and a necessary choice exist. If these conditions are satisfied, TRADE_OFF is assigned; otherwise, TENSION is assigned, provided that the opposition of tendencies has actually been established.

If statements are compatible and complement the description of the same phenomenon, COMPLEMENT is assigned. If no meaningful semantic relationship has been established and there is no specific uncertainty requiring clarification, UNRELATED may be assigned.

These rules should not be regarded as a universal hierarchy of classes. They constitute a procedure for selecting the primary label when semantic relationships potentially overlap.

For complex cases, a secondary label may also be assigned. However, this possibility should not lead to the arbitrary assignment of multiple classes. Each secondary label must be accompanied by a separate justification explaining the specific aspect in which the corresponding relationship also holds.

For example, a counterexample to a universal statement may receive COUNTEREXAMPLE as its primary label and CONTRADICTION as its secondary label. By contrast, simultaneously assigning TRADE_OFF and TENSION without explaining the underlying resource mechanism is insufficiently justified.

5.11. Abstention from Classification (NEEDS_CONTEXT)

In real-world information analysis tasks, it is not always possible to establish a semantic relationship between two statements unambiguously. This may result from undefined terminology, missing temporal characteristics, ambiguous quantifiers, unknown resource constraints, or insufficient information about the objects involved.

For such situations, SMF-MAP provides the auxiliary label NEEDS_CONTEXT.

It is not a tenth class of the taxonomy because it does not characterize an independent semantic relationship between statements. Its purpose is to indicate that the available data are insufficient to justify the selection of a primary label.

For example:

A: “The system operates quickly.”

B: “The system operates slowly.”

Without clarifying whether the statements refer to the same system, the same workload, the same time interval, and common performance criteria, it is impossible to determine whether the relationship is CONTRADICTION or CONTEXT_DEPENDENT.

In such a case, the procedure should not arbitrarily select a class but should indicate the need to clarify the conditions.

The practical significance of NEEDS_CONTEXT lies in reducing the number of unjustified classifications. At the same time, the frequency with which this label is used may serve as an independent indicator of corpus complexity or insufficiently specified rules.

Justified abstention is preferable to unjustified confidence.

5.12. Operational Procedure for Establishing Semantic Relationships

Applying the taxonomy in SMF-MAP involves a sequential examination of the meaning of statements. At the first stage, the objects, their properties, temporal characteristics, and domain context are identified. Next, quantifiers, negations, numerical thresholds, modal constructions, and other elements affecting the logical structure are determined.

The procedure then examines the possibility of bidirectional and unidirectional logical entailment, incompatibility of truth values, the presence of a specific counterexample, and contextual explanations for discrepancies.

If no corresponding formal logical relationship is established between the statements, the analysis proceeds to semantic connections: resource trade-offs, tension between tendencies, complementarity, or the absence of a significant relationship.

The generalized procedure can be represented as a mapping

$$f(A, B, K) \rightarrow (r, d, c),$$

where r is the primary semantic label or NEEDS_CONTEXT; d is the direction of the relationship, where applicable; and c is a reasoned explanation of the classification.

To improve traceability, it is also advisable to record alternative labels, identified ambiguities, required clarifications, and the results of critical verification.

A fundamental requirement is the prohibition of implicitly supplementing statements with conditions that do not follow from the input data. For example, if the statements do not specify memory constraints, the classifier must not independently assume that the total memory capacity is fixed and fully utilized merely to establish a TRADE_OFF relationship.

Similarly, the absence of an explicitly specified time interval does not justify automatically assuming that two statements refer to the same point in time.

In this way, the operational taxonomy combines the rules of logical analysis with requirements for contextual specificity and the justification of classification decisions.

5.13. Significance of the Taxonomy for Generating and Verifying Semantic Milestones

The proposed taxonomy performs several interrelated functions within SMF-MAP.

First, it enables the identification of semantic duplication. If two milestones are equivalent, the system may retain one as the primary formulation and the other as an alternative formulation.

Second, the taxonomy enables the identification of contradictory statements and counterexamples. This is particularly important for preventing mutually incompatible generalizations from being included in the results without appropriate explanations.

Third, logical entailment relationships make it possible to construct directed chains of statements that represent transitions from stronger statements to weaker ones or from general conditions to their consequences.

Fourth, the CONTEXT_DEPENDENT, TRADE_OFF, and TENSION classes provide a means of representing complex situations in which differences between statements cannot be reduced to simple contradiction or logical entailment.

Fifth, COMPLEMENT makes it possible to combine mutually complementary milestones into more complete semantic configurations, while UNRELATED allows statements between which no significant relationship has been established to be separated.

Thus, the taxonomy is used not only as a tool for classifying existing statements but also as a component of the procedure for generating them. The results of relationship analysis may provide grounds for refining, combining, rejecting, or regenerating semantic milestones.

At the same time, the operational specificity of the rules does not mean that their effectiveness has already been confirmed by an independent experiment. The single-session audit conducted within the study made it possible to identify ambiguous cases, clarify the direction of logical entailment, and detect several errors in the test examples. Evaluating classification accuracy under independent conditions requires separate model runs, expert-validated corpora, and controlled access to author-assigned labels.

Therefore, the nine-class operational taxonomy of SMF-MAP provides a formal basis for analyzing relationships between short, meaningful statements. Its use enables a transition from generating individual semantic milestones to constructing structured knowledge configurations that account for logical compatibility, context dependence, alternative explanations, and limits of applicability.

6. Materials and Research Protocol

6.1. Sequence of Corpus Development

The study proceeded in stages (Phases 1–18). The initial plan included 30 tasks of three types (TEXT, SITUATION, TOPIC), four conditions C0–C3, and three repetitions, amounting to 360 planned runs and up to 3,600 milestones. This plan was not implemented as actual independent model runs; the data presented below originate from a separate task involving the classification of pairs of statements.

First, a corpus of 90 pairs was constructed, with 10 pairs for each of the nine relationships. Following an editorial audit, version v1.1 of the rules was established, and a control set of 45 new author-constructed pairs was prepared (five per class). Subsequently, two modifications were created for each control pair: PRESERVE, intended to preserve the semantic relationship, and CHANGE, intended to alter it. This yielded 90 modifications in total. Following the initial evaluation, seven transitions were reviewed, the rules were refined to version v1.1, and simple linguistic frames were replaced with more structurally varied paraphrases.

6.2. Data Status and Reproducibility

All 90 primary pairs, 45 control pairs, and 90 modified pairs were created within the research process with the involvement of the same language model. They do not constitute a random sample of naturally occurring texts. The control set was not independent of the authoring model: despite the absence of exact duplicates, information about the classes and reference labels remained accessible in the context. Therefore, the term “control” describes the role of the corpus in the sequence of editorial checks rather than a fully blinded holdout set.

For some preliminary stages, deterministic simulations of agent roles and a plan for blinded tasks were developed, but simulated responses do not constitute independent judgments by experts or models. The Fleiss κ statistic derived from such a simulation was not included in the empirical results. All evaluations presented below were obtained without independent experts, without API-based orchestration of agent roles, and without experimental isolation of author-assigned labels.

6.3. Software Implementation and Open Repository

The author has made the full version of SMF-MAP, including framework materials and prompts for organizing agent roles, available in an open GitHub repository: <https://github.com/DmytroLande/SemanticMileStone/> [8]. The repository is identified as a source of software and methodological materials required to examine the implementation and subsequently reproduce the procedure. The description of its contents is based on information provided by the author; the repository's availability and current contents were not independently verified during the preparation of this version of the article.

The availability of published prompts and framework materials does not mean that all planned multi-agent runs have been performed or that the single-session evaluations reported in this article resulted from independent agent orchestration. Future studies should record the repository version (commit), model configurations, prompts, generation parameters, and execution logs to ensure reproducibility.

6.4. Metrics

For PRESERVE, SPR is defined as

$$SPR = \frac{N(P : L_{model} = L_{source})}{N_P}.$$

For CHANGE, SCDR is defined as

$$SCDR = \frac{N(C : L_{model} = L_{expected})}{N_C}.$$

FSR and FIR are calculated separately:

$$FSR = \frac{N(C : L_{model} = L_{source})}{N_C},$$

$$FIR = \frac{N(P : L_{model} \neq L_{expected})}{N_P}.$$

These are operational measures of label agreement. If an expected label proves to be logically incorrect, SCDR cannot be interpreted as the proportion of correct classifications. If a PRESERVE modification actually changes the meaning, FIR does not demonstrate a classifier error.

7. Results

The results presented in this section concern the development and internal verification of the operational taxonomy of semantic relationships used in SMF-MAP. The evaluation was organized as a sequence of corpus construction, preliminary classification, editorial revision, and controlled modification of statement pairs.

The analysis focused on whether the classification procedure could distinguish different semantic relationships, preserve classifications under meaning-preserving paraphrasing, and recognize changes introduced by modifications to logically significant elements of statements.

The evaluation involved three groups of materials: a primary corpus of 90 statement pairs, an additional corpus of 45 pairs, and 90 controlled modifications of the additional corpus. The latter comprised 45 meaning-preserving transformations and 45 transformations intended to change the semantic relationship.

The numerical results should be interpreted as outcomes of a sequential, single-session model-based audit. The same language model participated in corpus construction, reference-label assignment, and subsequent analysis, while information about expected classifications remained available in the shared context. Consequently, the reported agreement rates do not constitute independently measured classification accuracy or experimental confirmation of the effectiveness of the complete multi-agent architecture.

For traceability, test cases were assigned identifiers. The prefix R denotes a statement pair from the primary corpus, while H denotes a pair from the additional corpus. The numerical component identifies the individual pair. For modified examples, the suffix P indicates a transformation intended to preserve the semantic relationship (PRESERVE), whereas C indicates a transformation intended to change it (CHANGE). Thus, H023-P refers to a meaning-preserving modification of pair 23 in the additional corpus. These identifiers serve to connect the analysis with the underlying test materials and are not themselves semantic categories.

The evaluation results are presented below, together with an analysis of the main sources of classification ambiguity and the methodological implications of the observed disagreements.

7.1. Primary Corpus: 90 Pairs

The first stage involved the construction of a primary corpus containing 90 pairs of statements. The corpus was designed to cover the nine semantic relationship classes defined in Section 5: CONTRADICTION, COUNTEREXAMPLE, EQUIVALENCE, ENTAILMENT, CONTEXT_DEPENDENT, TRADE_OFF, TENSION, COMPLEMENT, and UNRELATED.

Each class was represented by ten pairs. This balanced structure was intended to facilitate the examination of classification rules and to prevent the analysis from concentrating disproportionately on the most obvious logical relationships.

In the initial single-session classification, 87 of the 90 assigned primary labels agreed with the preliminary reference labels, corresponding to 96.7% agreement. Three disagreements were recorded.

The first disagreement concerned the distinction between TENSION and TRADE_OFF (R049). These relationships may appear similar because both involve competing objectives or opposing tendencies. However, TRADE_OFF requires an identifiable limitation or shared resource that makes the simultaneous satisfaction of competing objectives impossible or constrained. TENSION is a broader relationship in which tendencies or objectives oppose one another without a demonstrated necessary resource trade-off.

The second disagreement concerned CONTRADICTION and COUNTEREXAMPLE (R003, recorded as R03 in the initial protocol). A specific observation that refutes a universal statement may also establish a logical contradiction. Nevertheless, the operational taxonomy treats COUNTEREXAMPLE as the primary relationship when the structure of the pair explicitly involves a general proposition and a particular instance that invalidates it. This distinction illustrates why a single pair may support both a primary and a secondary semantic relationship.

The third disagreement concerned TENSION and COMPLEMENT (R050). Two statements may describe different consequences of the same process without being logically incompatible. For example, one statement may describe an improvement in performance, while another describes increased complexity or reduced transparency. Whether such statements should be classified as complementary or as expressing tension depends on the relationship between the corresponding effects and the context in which they are evaluated.

These disagreements were not automatically interpreted as classification errors. Instead, they prompted an editorial examination of the reference examples and the criteria used to distinguish neighboring semantic classes.

The editorial audit produced four categories of decisions. Thirty-eight pairs were retained without modification. For 41 pairs, a secondary semantic relationship was permitted because the primary label did not exhaust all meaningful aspects of the relationship. Ten pairs were designated for rewriting to eliminate ambiguity or insufficiently specified conditions. In one case, the reference label itself was corrected.

Following 12 editorial modifications, a balanced corpus of 90 pairs was prepared, with ten primary labels for each semantic class.

The principal outcome of this stage was therefore not the numerical agreement rate alone. More importantly, the analysis revealed that semantic relationships cannot always be treated as mutually exclusive categories without specifying the rules for selecting the primary relationship.

In particular, the audit demonstrated the need to distinguish a direct contradiction from a counterexample, a resource-dependent trade-off from a general tension, and complementarity from opposing tendencies. These distinctions were incorporated into the refinement of the operational taxonomy.

The 96.7% agreement observed in the initial pass should not be interpreted as an independent estimate of classifier accuracy. The corpus was constructed specifically for the study, and the classification procedure had access to the reference-label context. The result instead characterizes the consistency of the preliminary examples and rules within the internal audit.

7.2. Additional Corpus: 45 Pairs

To examine the refined classification rules on additional material, a second corpus of 45 statement pairs was constructed. It contained five examples for each of the nine semantic relationship classes.

The purpose of this corpus was to extend the range of examples and examine whether the operational distinctions established during the primary audit could also be applied to newly constructed statement pairs.

The additional corpus did not constitute an independently sampled or blinded test set. Although its examples were different from those in the primary corpus, they were created within the same research process, and information about the classification rules and expected labels remained available to the model.

In the single-session analysis, the primary classifications agreed with all 45 author-assigned reference labels. This corresponds to complete agreement within the specified evaluation procedure.

However, the subsequent examination showed that agreement with the primary label did not necessarily mean that the corresponding example was unambiguous or factually correct.

Secondary semantic relationships were identified for 11 pairs. This indicates that some examples contained additional meaningful relationships beyond their assigned primary categories. Such cases are particularly important for the proposed taxonomy because they demonstrate that semantic classification may require both a primary decision and an accompanying explanation of alternative interpretations.

Substantive comments were also recorded for 11 pairs. These comments concerned the formulation of statements, implicit assumptions, factual details, and the conditions required to justify particular relationships.

One illustrative case was H007. The example contained a statement referring to the expression "Key found" as consisting of three words. In fact, the expression contains two words. The inconsistency did not necessarily invalidate the assigned COUNTEREXAMPLE relationship, but it revealed an error in the factual content of the test material.

This distinction is methodologically important. A model may correctly identify the logical relationship between two statements while failing to detect an error in the statements themselves. Conversely, a factual error may exist without changing the formal relationship assigned to a pair.

The analysis therefore confirmed the need to separate at least two verification tasks: determining the semantic relationship between statements and examining the correctness of the information contained in those statements.

The additional corpus also provided the basis for the subsequent modification experiments. Each of its 45 pairs was used to construct two modified versions, allowing the study to examine both the preservation and alteration of semantic relationships.

Thus, the complete agreement observed for the additional corpus should be regarded as an intermediate result of the internal audit rather than evidence of perfect classification accuracy. The substantive comments and secondary relationships were at least as informative as the agreement rate because they revealed limitations that would remain invisible in a simple comparison of primary labels.

7.3. First Series of Modifications: 90 Pairs

The next stage examined how the semantic classification responded to controlled modifications of the additional corpus.

For each of the 45 original pairs, two modified versions were constructed. The first modification, designated PRESERVE, was intended to change the linguistic formulation without changing the original semantic relationship. The second, designated CHANGE, was intended to alter a logically significant element and thereby produce a different relationship.

The resulting evaluation set contained 90 modifications: 45 PRESERVE and 45 CHANGE cases.

The distinction between these transformations is essential. A meaning-preserving modification may replace words with synonyms, change sentence structure, or express the same numerical value in another form. A meaning-changing modification may alter a quantifier, introduce or remove negation, change a numerical threshold, modify temporal conditions, or reverse the direction of logical entailment.

For example, replacing "80 milliseconds" with "0.08 seconds" preserves the numerical value. Provided that the remaining conditions are unchanged, such a transformation should not alter the semantic relationship.

By contrast, replacing "all agents completed the task" with "at least one agent completed the task" changes the logical strength of the proposition. The two formulations are not equivalent, even though they concern the same agents and task.

These examples illustrate the intended distinction between the two transformation types; they should not be interpreted as a complete reproduction of the corresponding test records.

Four operational metrics were used to characterize the results.

Semantic Preservation Rate (SPR) measures the proportion of PRESERVE modifications for which the original relationship label was retained.

Semantic Change Detection Rate (SCDR) measures the proportion of CHANGE modifications for which the assigned label exactly matched the expected label.

False Stability Rate (FSR) measures the proportion of CHANGE modifications for which the original label was retained despite an intended semantic change.

False Instability Rate (FIR) measures the proportion of PRESERVE modifications for which the assigned label differed from the expected preserved label.

These metrics describe label agreement under the specified transformations. Their interpretation depends on the validity of the expected labels and on whether the modifications actually preserve or change the intended meanings.

In the first modification series, all 45 PRESERVE cases retained their original classifications, producing $SPR = 45/45$, or 100.0%. No false instability was recorded, giving $FIR = 0/45$.

For the 45 CHANGE cases, exact agreement with the expected labels was obtained in 38 cases, corresponding to $SCDR = 38/45$, or 84.4%. The original relationship label was not retained in any CHANGE case, giving $FSR = 0/45$.

Seven CHANGE cases produced labels different from the expected reference labels: H011-C, H023-C, H026-C, H031-C, H035-C, H040-C, and H045-C.

The examination of these disagreements identified several recurring methodological problems. Some concerned the direction of logical entailment, which must be explicitly specified when one statement follows from another but the reverse implication does not hold. Others concerned the role of resource constraints in distinguishing TRADE_OFF from TENSION. Additional difficulties arose from numerical thresholds, modal expressions, and implicit contextual assumptions.

These disagreements demonstrated that recognizing a change in meaning and identifying the exact resulting semantic relationship are different tasks. A classification procedure may detect that a statement has changed without assigning the particular new relationship expected by the author.

The perfect preservation result also required cautious interpretation. The initial PRESERVE modifications predominantly involved relatively simple linguistic transformations and neutral changes in formulation. Such modifications did not sufficiently challenge the classification rules because they often left the logical structure of the statements almost unchanged.

Consequently, the first modification series was useful for identifying weaknesses in the operational criteria but provided limited evidence concerning robustness under more substantial linguistic restructuring.

This observation motivated a subsequent revision of the transformation procedure, with greater emphasis on structural paraphrasing and logically sensitive expressions.

7.4. Enhanced Series: Structural Paraphrasing

The enhanced evaluation series was designed to make the meaning-preserving transformations more demanding and to examine whether the classification rules remained stable when the linguistic structure of statements changed substantially.

Unlike the initial series, which relied mainly on simple paraphrasing, the revised PRESERVE transformations included component reordering, synonymous constructions, unit conversions, and changes in the syntactic representation of quantifiers.

For example, the numerical expression "80 milliseconds" was replaced with "0.08 seconds." Similarly, "less than 100 milliseconds" was reformulated as "did not reach one-tenth of a second."

These formulations are numerically equivalent when they refer to the same measurement and use identical boundary conditions. Nevertheless, they differ in linguistic structure and may require additional reasoning to establish that the semantic relationship remains unchanged.

Such transformations are especially relevant to semantic milestone analysis because short statements frequently express quantitative conditions indirectly. A minor change in wording may preserve meaning, introduce ambiguity, or alter the logical relationship, depending on the interpretation of the terms used.

The enhanced evaluation also examined modifications involving quantifiers and technical terminology. These cases are more difficult than ordinary synonym replacement because semantic equivalence depends on logical scope, accepted definitions, and contextual assumptions.

The results of the initial and enhanced modification series are summarized in Table 2.

Table 2. Comparison of Semantic Classification Results in the Initial and Enhanced Modification Series

Metric	Initial series (Phase 16)	Enhanced series (Phase 18)
Semantic Preservation Rate (SPR)	45/45 = 100.0%	44/45 = 97.8%
Semantic Change Detection Rate (SCDR), exact label agreement	38/45 = 84.4%	36/45 = 80.0%
False Stability Rate (FSR)	0/45 = 0.0%	0/45 = 0.0%
False Instability Rate (FIR)	0/45 = 0.0%	1/45 = 2.2%

Note: Phase 16 and Phase 18 denote stages of the sequential research audit. They do not represent independent experimental replications. The two series used revised transformations and reference conditions, and their results should not be interpreted as a controlled comparison of an unchanged model on identical data.

In the enhanced series, 44 of the 45 meaning-preserving modifications retained the original relationship label, corresponding to $SPR = 97.8\%$. One PRESERVE case produced a different label, giving $FIR = 2.2\%$.

For meaning-changing modifications, exact agreement with the expected label was obtained in 36 of 45 cases, corresponding to $SCDR = 80.0\%$. The False Stability Rate remained zero, indicating that none of the CHANGE cases retained the original label under the operational definition used in the audit.

The decrease in SPR from 100.0% to 97.8% indicates that the revised PRESERVE transformations introduced at least one case in which the classification did not remain stable. However, this observation does not by itself establish that the model made an error. The disagreement may also reflect an insufficiently specified reference label or a transformation that did not fully preserve the intended meaning.

Similarly, the decrease in exact $SCDR$ agreement from 84.4% to 80.0% should not be interpreted as evidence of deteriorating classifier performance. The enhanced series contained more demanding transformations, and the reference conditions were revised during the research process.

The two series therefore cannot support a direct quantitative comparison of performance under identical experimental conditions.

Their combined methodological significance lies elsewhere. They demonstrate that the apparent stability of semantic classification depends strongly on the types of transformations used in evaluation. Simple paraphrases may produce high agreement while failing to reveal weaknesses associated with quantification, technical definitions, numerical conditions, and contextual assumptions.

The enhanced series also showed the importance of distinguishing between a change in the assigned label and a change in the underlying meaning. A label disagreement is informative only when the reference interpretation has been independently justified.

The observed results therefore support the use of more structurally diverse transformations in future evaluation protocols, together with explicit documentation of the assumptions under which each modification is expected to preserve or change the semantic relationship.

7.5. Analysis of Ambiguous Semantic Transitions

The detailed examination of the enhanced evaluation series revealed several recurring sources of ambiguity. These were associated primarily with quantification, the relationship between universal and individual statements, numerical thresholds, graph-theoretic terminology, and the distinction between related methodological concepts.

The disagreements were particularly informative because they exposed cases in which a seemingly minor linguistic change altered the logical interpretation of a statement or revealed an insufficiently specified reference label.

For clarity, Table 3 summarizes the main problematic cases. The identifiers correspond to the test records described in Section 7, while the explanations describe the semantic issues responsible for the observed ambiguities.

Table 3. Main Sources of Ambiguity Identified in the Enhanced Semantic Classification Audit

Test case	Semantic issue	Explanation and interpretation
H001-C	Numerical quantification	"Exactly five" excludes additional instances, whereas "there are five" may not. The two formulations are not necessarily equivalent.
H006-C, H008-C, H009-C, H010-C	Universal statements and individual examples	Relationships between general propositions and specific instances depend on set membership, existence assumptions, and the scope of quantification.
H007-C	Disagreement requiring further examination	This case was among the Phase 18 disagreements, but the available summary does not establish its exact logical cause.
H015-C	Definition of a graph property	Equivalence depends on the precise formal definition of the graph-theoretic property under consideration.
H023-P	Qualitative versus quantitative conditions	Descriptions such as "low latency" do not define numerical thresholds unless the context provides them.
H035-C	Openness and reproducibility	Publishing a methodology, making it openly accessible, and enabling independent reproduction are related but distinct properties.
H041-C	Graph-theoretic terminology	The meaning of "closed path" depends on the adopted definition and must be distinguished from related concepts such as a closed walk or cycle.

The table identifies the main semantic problems without assuming that every disagreement represents an incorrect model decision. In several cases, the available information was insufficient to establish whether the expected label itself was justified.

Quantification and Logical Strength

One of the most important sources of ambiguity concerned numerical quantifiers.

The difference between "exactly five" and "there are five" illustrates the problem. The first formulation explicitly specifies the number of relevant objects and excludes additional instances. The second may be interpreted as indicating the existence of five objects without necessarily excluding others.

For example, the statement "Exactly five agents completed the task" establishes a precise numerical condition. The statement "Five agents completed the task" may express the same condition in ordinary language, but it may also be interpreted as asserting that at least five agents completed the task.

The semantic relationship between these formulations therefore depends on the accepted interpretation of the numerical expression.

This problem was identified in H001-C and demonstrates why numerical similarity cannot automatically be treated as logical equivalence.

Universal Statements and Specific Instances

A second group of disagreements concerned the relationship between universal propositions and individual observations.

Consider a general statement asserting that all elements of a particular set possess a property. Such a statement may entail that a specific element possesses the same property, but only if the element belongs to the set and the necessary existence conditions are satisfied.

Conversely, observing that one element possesses a property does not justify the conclusion that every element of the set possesses it.

The problem becomes more complicated when the statements involve negation or counterexamples. A single verified instance may refute a universal statement, but only when the instance falls within the domain covered by that statement.

Cases H006-C, H008-C, H009-C, and H010-C illustrated the importance of these conditions.

The corresponding disagreements showed that the classification of ENTAILMENT, CONTRADICTION, and COUNTEREXAMPLE requires explicit consideration of quantifier scope, set membership, and the direction of logical implication.

Ambiguity of Technical Terminology

Another source of disagreement concerned graph-theoretic concepts.

Technical terms may appear semantically equivalent in ordinary language while having different meanings under formal definitions.

For example, the terms "path," "walk," and "cycle" describe related but distinct graph structures. Their interpretation depends on whether repeated vertices or edges are permitted and whether the initial and final vertices coincide.

The expression "closed path" is particularly sensitive to the adopted terminology. In some contexts, it may be used informally to describe a route returning to its starting vertex, while formal graph theory distinguishes several types of closed structures.

Case H041-C demonstrated that semantic classification requires an explicit definition of the graph-theoretic object under consideration.

A related difficulty occurred in H015-C, where the expected equivalence depended on the precise definition of a graph property.

These cases indicate that the correct classification of technical statements may require access to domain-specific definitions rather than relying exclusively on linguistic similarity.

Qualitative and Quantitative Conditions

Case H023-P concerned a transformation intended to preserve the semantic relationship but involving qualitative descriptions of latency.

Expressions such as "low latency," "fast response," or "short processing time" are inherently dependent on the application context. Without numerical thresholds, they do not define precise measurable conditions.

For example, a response time of 80 milliseconds may be considered low in one application but unacceptable in another. Consequently, replacing a numerical condition with a qualitative description does not necessarily preserve the original meaning.

This problem is especially important for semantic milestones because concise formulations often omit numerical details.

A meaning-preserving transformation must retain not only the general topic of the statement but also the conditions under which its truth value is determined.

Openness, Publication, and Reproducibility

Case H035-C revealed a distinction between several methodological properties that are frequently treated as equivalent.

The publication of a methodology makes information about a procedure available to readers. Openness may additionally concern access to software, data, prompts, or other research materials. Reproducibility requires that the available information and resources be sufficient for an independent researcher to repeat the procedure under specified conditions.

These properties are related, but none can automatically be substituted for the others.

For example, publishing a description of an algorithm does not guarantee that its results can be reproduced if essential data, model configurations, or implementation details are unavailable.

Similarly, making source code publicly accessible does not necessarily establish reproducibility when the experimental environment and evaluation procedure are insufficiently documented.

The ambiguity of H035-C remained unresolved after the revision of the corresponding transition.

This result illustrates the importance of distinguishing conceptual similarity from logical equivalence, particularly when statements concern methodological requirements.

Interpretation of Disagreements and Contextual Abstention

In the enhanced evaluation series, ten modifications produced classifications that differed from their expected reference labels: H001-C, H006-C, H007-C, H008-C, H009-C, H010-C, H015-C, H023-P, H035-C, and H041-C.

Seven modifications received the NEEDS_CONTEXT designation.

Unlike the nine semantic relationship classes, NEEDS_CONTEXT does not describe a relationship between two statements. It indicates that the available contextual information is insufficient to establish a justified classification.

This distinction is important because the absence of a clearly established relationship is not equivalent to the absence of any meaningful relationship.

For example, two statements may appear contradictory until their temporal conditions are clarified. Similarly, two numerical statements may appear equivalent until the interpretation of their quantifiers is examined.

Assigning UNRELATED in such situations would be misleading because the problem is not necessarily the absence of a relationship but the insufficiency of information required to identify it.

The NEEDS_CONTEXT designation therefore provides a mechanism for avoiding unjustified classification decisions.

Of the seven specifically revised transitions examined during the subsequent semantic verification, six passed the verification procedure, while H035-C remained ambiguous.

These results demonstrate that disagreements with expected labels may have different explanations. Some may reflect incorrect classification decisions, others may result from insufficiently specified statements, and still others may reveal weaknesses in the author-assigned reference labels.

Accordingly, the evaluation of semantic classification should not rely exclusively on numerical agreement with predefined labels. It should also examine whether the labels are logically justified, whether the context is sufficiently specified, and whether alternative interpretations have been considered.

The principal outcome of this analysis is the recognition that the verification of semantic relationships and the verification of reference classifications are closely connected but distinct tasks.

A classification procedure may disagree with an expected label precisely because it identifies an ambiguity that the original annotation overlooked.

For this reason, the internal audit should be regarded primarily as a means of refining the operational taxonomy, identifying problematic examples, and clarifying the conditions under which semantic relationships can be established.

The results provide a methodological basis for future independent evaluation, but they do not establish the generalization accuracy of SMF-MAP or the superiority of its multi-agent architecture over simpler approaches.

The analysis also demonstrates a broader principle underlying the proposed framework: reliable semantic classification requires not only the recognition of relationships between statements but also the ability to explain the conditions under which those relationships hold, identify insufficient evidence, and refrain from unsupported conclusions.

8. Discussion and Conclusions

The study made it possible to formulate the conceptual foundations of SMF-MAP (Semantic Milestone Framework – Multi-Agent Pipeline), define the structure of a semantic milestone, develop an operational taxonomy of relationships between statements, and conduct a staged audit of the rules for their classification. The results indicate that the generation of short, meaningful statements should be regarded not as an isolated text generation task but as a comprehensive procedure for representing, comparing, and verifying knowledge. At the same time, the findings require a clear distinction between the proposed theoretical principles, the results of the internal model-based audit, and the characteristics that can be established only through independent experimentation.

One of the findings of the study was the identification of three interrelated levels of semantic reliability. The first level concerns the local correctness of an individual statement. Even a semantically valid milestone may contain errors in its formal characteristics, particularly in word counts, quantitative parameters, or descriptions of applicability conditions. Such errors do not necessarily alter the essential meaning of a statement, but they reduce the reproducibility of the procedure and confidence in its results. Therefore, formal verification should be separated from semantic evaluation and performed according to unambiguously defined rules.

The second level of semantic reliability concerns the correctness of establishing relationships between statements. The staged audit confirmed the importance of preserving object identity, temporal characteristics, quantifiers, the scope of negation, numerical thresholds, and contextual conditions. The absence of even one of these specifications may change the classification outcome. For example, two statements that appear contradictory may refer to different points in time, while statements with similar linguistic structures may differ in the direction of logical entailment. Consequently, semantic comparison cannot be limited to determining lexical similarity or vector-based proximity between texts.

The third level concerns the quality of the reference data used to verify classification decisions. The audit revealed cases in which the original author-assigned label required clarification or correction. This finding has fundamental methodological significance: agreement between a model's output and a reference label does not necessarily demonstrate the correctness of semantic

analysis if the reference label itself is erroneous. Conversely, a reasoned departure from the original label may represent a justified decision rather than a classifier error.

Agreement with a reference label does not guarantee the correctness of a conclusion.

This problem is particularly important in studies where large language models are involved simultaneously in generating test examples, assigning reference labels, and evaluating the resulting outputs. In such cases, there is a risk of interdependence among judgments: the model may reproduce previously established assumptions without independently verifying their validity. Therefore, the quality of the reference corpus should be treated as a separate object of quality control rather than as an inherently guaranteed property of the input data.

Boundary cases between individual classes of semantic relationships proved particularly informative for refining the operational taxonomy. Specifically, distinguishing CONTRADICTION from COUNTEREXAMPLE requires consideration of the logical structure of statements. A specific counterexample may simultaneously create an incompatibility with a universal statement; however, its primary function is precisely to refute the generalization. Therefore, COUNTEREXAMPLE should be assigned as the primary relationship when a specific case falls within the scope of a universal statement and refutes it.

Distinguishing TRADE_OFF from TENSION is equally important. Opposing tendencies or competing objectives do not always constitute a resource trade-off. Establishing TRADE_OFF requires identifying a shared limited resource and demonstrating that improving one outcome under the specified conditions necessarily restricts another. If no such mechanism has been established, but the statements reflect opposing tendencies, TENSION is the more justified classification. This distinction has practical significance for analyzing systems in which competing requirements may be reconciled through architectural changes, redistribution of functions, or allocation of additional resources.

For EQUIVALENCE, bidirectional verification of logical entailment is a necessary condition. Similarity of wording or overlap between individual concepts cannot substitute for such verification. At the same time, ENTAILMENT requires explicit preservation of the direction of entailment, since reversing it may lead to an unjustified strengthening of a conclusion. Accordingly, the refined SMF-MAP v1.1 rules introduce the subclasses ENTAILMENT_A_TO_B and ENTAILMENT_B_TO_A, which indicate the direction of a single basic semantic relationship.

Another outcome was the clarification of the status of the NEEDS_CONTEXT label. Its use makes it possible to distinguish cases in which no significant relationship has been established between statements from situations in which the available information is insufficient for a justified classification. This distinction prevents ambiguous pairs from being forcibly assigned to one of the nine basic classes and allows for subsequent contextual clarification.

Uncertainty about a relationship does not imply the absence of that relationship.

Taken together, these findings demonstrate that the operational taxonomy should perform not only a classification function but also a verification function. It should make it possible to explain why a particular relationship has been established between two statements, which conditions were

considered, which alternative classes were examined, and which circumstances might change the decision. This constitutes one of the conceptual distinctions between SMF-MAP and approaches primarily oriented toward generating short statements or determining their semantic similarity.

The study also considered the historical and philosophical aspects of formulating short, meaningful generalizations. An earlier version of the work compared six interpretations of motifs from Omar Khayyam, six motifs from the Tao Te Ching, and six demonstrative SMF semantic milestones. The purpose of this comparison was to illustrate different ways of formulating concise, self-contained, and semantically rich statements. The resulting mean scores of 8.80, 9.12, and 9.25, respectively, reflect the subjective evaluation of specially selected examples by a single language model. These evaluations were not subjected to independent blinded assessment, were not accompanied by comprehensive verification of the authenticity of the translations used, and were not included in the quantitative metrics of the final audit phase. Therefore, this comparison has exclusively conceptual and illustrative significance and cannot be regarded as evidence of the superiority of SMF-MAP over historical or literary approaches to formulating aphoristic statements.

The quantitative component of the study comprised 90 primary pairs of statements, 45 additional control pairs, and 90 modifications, of which 45 were intended to preserve the original semantic relationship and 45 to change it. This organization of the material made it possible to examine the behavior of classification rules under paraphrasing, contextual clarification, and modifications to logically significant elements of statements.

In the enhanced audit series, the Semantic Preservation Rate (SPR) was 97.8%, while the Semantic Change Detection Rate (SCDR) was 80.0%. The first metric represents the proportion of cases in which the original classification was preserved under semantically equivalent transformations, while the second represents the proportion of relationship changes correctly identified within the corresponding evaluation series. However, these values were obtained within a single-session model-based audit and therefore cannot be interpreted as independent estimates of SMF-MAP classification accuracy, the robustness of the multi-agent architecture, or the system's ability to generalize to unseen corpora.

The principal methodological limitation of the study is the lack of independence between corpus construction, reference-label assignment, and result verification. The use of a shared model context creates the possibility of transferring previous decisions between stages and therefore prevents the observed agreements from being treated as independent observations. Additional limitations include the relatively small corpus size, manual balancing across classes, the simplified nature of some examples, ambiguity in certain terms, and editorial modifications to the test material between audit phases.

Another substantial limitation is the absence of isolated runs of individual agents and a comprehensive experimental comparison of configurations C0–C3. These configurations were included in the research plan, but their independent execution with the required number of repetitions was not performed. Consequently, the available results do not permit a quantitative determination of the contribution of individual agent roles, an assessment of the advantages of multi-agent organization over single-stage generation, or an evaluation of the statistical significance of such advantages.

Furthermore, metrics obtained at different stages of corpus editing and rule refinement cannot be mechanically combined into a single accuracy estimate. Changes in the composition of examples, corrections to reference labels, and modifications to classification criteria mean that the results of individual phases characterize evaluation conditions that are not entirely identical. They are suitable for analyzing the development and refinement of the taxonomy but do not constitute a single independent experimental dataset.

Thus, the audit should primarily be regarded as a methodological examination of semantic analysis rules, the identification of ambiguities, and quality control of reference examples. Its significance lies in documenting problems that may remain undetected when model outputs are formally compared with their original labels.

Further empirical validation of SMF-MAP requires the construction of a predefined and fixed corpus, independent annotation by at least several experts, an established procedure for resolving disagreements between annotators, and isolation of reference labels from the models performing classification. It is also advisable to conduct independent repeated runs for configurations C0–C3, preserve request and response logs, record model versions, and statistically evaluate the consistency of the results.

Particular attention in future experiments should be given to boundary cases within the taxonomy, as these make it possible to evaluate not only the system's ability to reproduce obvious examples but also the correctness of its application of operational criteria. Another research direction should involve evaluating the contributions of the Critic Agent, Verifier Agent, and Semantic Comparator to the quality of final decisions. This requires controlled ablation experiments in which individual functional components are systematically removed or replaced while all other conditions remain unchanged.

Another promising research direction is the transition from classifying individual pairs of statements to analyzing complete semantic configurations. Such configurations may simultaneously contain relationships of equivalence, logical entailment, contradiction, context dependence, and complementarity. Their joint analysis creates opportunities for detecting hidden inconsistencies, exploring alternative explanations, and constructing semantically coherent collections of milestones.

As a result of the study, a formal model of a semantic milestone was proposed as a structured object combining a short statement with its functional type, context, supporting evidence, quality characteristics, possible consequences, and temporal parameters. This model makes it possible to preserve the conciseness of the external formulation without discarding information required for verification and interpretation.

A conceptual multi-agent architecture, SMF-MAP, was developed to integrate the generation of candidate milestones, semantic analysis, critical examination, verification, editorial synthesis, and coordination of agent procedures. The architecture provides for a shared information base, feedback mechanisms, decision logging, and human involvement in evaluating the results.

The operational taxonomy of semantic relationships was refined to include nine basic classes: CONTRADICTION, COUNTEREXAMPLE, EQUIVALENCE, ENTAILMENT, CONTEXT_DEPENDENT, TRADE_OFF, TENSION, COMPLEMENT, and UNRELATED. For

ENTAILMENT, explicit specification of the direction of logical entailment was introduced, while NEEDS_CONTEXT was defined as the outcome of justified abstention from classification when contextual information is insufficient.

The staged single-session audit made it possible to identify errors and ambiguities in reference examples, refine the rules for distinguishing related classes, and examine how classification responds to controlled changes in wording. The resulting quantitative metrics characterize the outcomes of this internal audit but do not establish the independent effectiveness of the multi-agent architecture.

Therefore, the principal scientific contribution of this work lies in formalizing the procedure for generating and semantically validating short, meaningful statements, defining operational criteria for comparing them, and documenting the conditions required for subsequent experimental verification. The proposed approach provides a methodological foundation for moving from the generation of individual aphoristic statements toward the construction of structured and verifiable knowledge configurations.

References

- [1] Nenkova, A. and McKeown, K., 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3), pp.103-233.
- [2] Reimers, N. and Gurevych, I., 2019, November. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992).
- [3] Bowman, S.R., Angeli, G., Potts, C. and Manning, C.D., 2015, September. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632-642).
- [4] Dung, P.M., 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2), pp.321-357.
- [5] Khayyam, O. and FitzGerald, E., 2010. *Rubáiyát of Omar Khayyám*. Oxford University Press.
- [6] Laozi, 1995. *The Tao Teh King, Or, The Tao and Its Characteristics*. Project Gutenberg.
- [7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, pp.9459-9474.
- [8] Lande D. SemanticMileStone. GitHub. URL: <https://github.com/DmytroLande/SemanticMileStone/>