Direct Semantic Reasoning Unit: The O(1) AI Primitive That Reasons In Latent Space

Executive Summary

I have discovered how to make neural networks perform complex reasoning in constant time, regardless of input complexity. The Direct Semantic Reasoning Unit (DSRU) processes entire thoughts, concepts, and tasks in a single forward pass—no tokens, no attention, no scaling bottlenecks.

Note that this occurs in latent space, and encoding and decoding are necessary to actually make use of this - which does add some additional overhead, but even in a very simplistic use case that doesn’t take advantage of any chaining, this is still enough to create a 93x throughput advantage in my example implementation.

This is achieved by direct semantic vector to semantic vector transformation, trained through supervised learning on various reasoning and instruct tasks, among others. The results were noteworthy:

1. A 93x throughput advantage over Zephyr 7B
2. A 19x latency advantage over Zephyr 7B
3. Identical accuracy running in batched mode to Zephyr 7B (77.7%) - unbatched accuracy for Zephyr was 80%
4. Core model inference time of ~1ms, end to end weighted GPU inference time per example across the full funnel (encoding, DSRU reasoning, decoding) of ~1.3ms

Reasoning From First Principles:

In an attempt to achieve faster task-specific inference, I began to evaluate existing ML architectures and tools and asked myself, "Can the elements of this be reconfigured in a way that enables a promptable, smart classifier?"

I ruled out many potential options (couldn't use softmax or attention because of scaling costs), but one that stuck with me was the idea of a direct vector to vector transformation. This was my reasoning:

1. Neural Nets are universal function approximators
2. Semantic embeddings are just vectors
3. Vector transformations can generally be achieved with functions
4. In principle, a neural network of sufficient depth and expressivity should be able to represent any such transformation or set thereof, provided there is enough information to enable the desired transformation

The core remaining question, from my perspective, was whether or not there was enough information in the semantic embedding to make reasoning or at least task completion possible. This went against the conventional wisdom of the field (not something I knew at the time), but the experiment was cheap and accessible - within a few days of experimentation, I found an approach that produced significant convergence on the NIV2 dataset, with validation accuracy steadily climbing. It plateaued earlier than one would need from a practical tool, but it showed clear gain in signal, even in a fairly naive attempt at implementation and a small model. From there, it was simply a matter of experimentation and engineering.

I've since come to believe that this is possible because modern semantic embeddings live in a sort of "inherently attended space". Semantic vector embeddings are essentially just the series of weighted sums that is output from the attention layer as an input to the hidden layers, with a set of transformations applied. This includes the terminal output of a semantic vector. If this is the case, applying attention to them isn't strictly necessary - they *are* the product of attention, albeit intentionally altered and compressed.

However, this is true of any new layer in a NN – it is a transformed (and potentially compressed or expanded) representation of what came before it – and for that reason, semantic embeddings are no different from a compressed version of the layer that came before it - projecting down to 1024 dimensions in the final layer is no different from projecting down to 1024 dimensions on any intermediate layer.

The unique part of semantic embeddings, however, is what they target as their outputs - in particular in the case of bge-large, a representation of meaning which is normalized to a unit hypersphere, making it trivial to compare to other meanings. However, it *appears* that despite being induced to conform to this format, it still has sufficient artifacts of attention to behave as an attended value.

This also implies that any neural network can follow a similar pattern, until it can't - outputting even one layer early as a vector and examining its output essentially gives you a "semantic embedding" in the sense that's a vector which has some sort of logical or linguistic meaning. What makes the modern semantic embeddings from a trained embedding engine *useful* is that unique ability to make simple mathematical comparisons to know their degree of relatedness in meaning.

This allows a primitive form of observability, even without advanced or computationally expensive techniques like Vec2Text – 'guess and check' works. While not sophisticated, guess and check is both fast and useful, when applied cleverly. It's entirely possible that we can create *more* sophisticated guess and check systems that allow for quick conclusion about what is in an embedding by performing a NNS against a set of summaries of various topics, precise enough to hone in on a meaning, but vague enough to allow for flexibility in interpretation. This cannot tell us what the embedding *says*…but it can tell us what it is *about*.

The DSRU is, essentially, the first *natively attended* model architecture. It doesn't provide its own attention; it borrows it from upstream elements in the system. Given the quadratic complexity of attention being applied on every forward pass, and the typical limitation of the transformer architecture to token-by-token operation, this essentially provides two dimensions of efficiency over a standard transformer; the elimination of the quadratic cost of attention in each forward pass, and expansion of the scope of the forward pass to include include the entire semantic embedding rather than operating over a single token.

Key Innovation: A neural architecture that performs semantic transformations in O(1) time, enabling reasoning over an entire 'thought' in each forward pass, rather than tokens or other linguistic subunits.

Proof of Concept: I demonstrate this capability through a promptable, intelligent classification system with inference-time configurable vocabulary that:

- Achieves 93x higher throughput than a Zephyr 7B LLM achieving comparable accuracy running on the same hardware
- Achieves 77.7% average accuracy across 13 tasks
- Performs all reasoning in a single forward pass of the integrated DSRU

Note that this is not what a DSRU *is* – this is merely an easily accessible (in fact, near-trivial) application of a DSRU.

The Paradigm Shift: Every AI system today is bottlenecked by sequential token processing. DSRU breaks this bottleneck, enabling real-time reasoning at scales previously thought impossible.

Patent Pending: The core pathways to train, integrate, and operate these models are subject to provision patent protection, and will become actual protection effective dated to early July 2025, provided I complete filing within 12 months.

⚠️ Technical Note on O(1) Complexity

When we refer to O(1) operation, we specifically mean the DSRU's core reasoning computation is constant-time regardless of the original input complexity. The complete flow in our example application outlined above includes:

- **Encoding (O(n))**: Converting variable-length text to fixed-size embeddings (~12ms, scales with input length)

- **DSRU Reasoning (O(1))**: Semantic transformation in constant time (~1ms, independent of input complexity)
- **Decoding (O(1))**: Mapping output embeddings to results (~0.1ms)

A helpful analogy: Think of DSRU like a hash table. Computing the hash from an input string is O(n), but the actual lookup/insert operation is O(1). We still consider hash tables to have O(1) operations even though we need to process the key first. Similarly, DSRU performs O(1) semantic transformations even though encoding the input takes O(n) time.

Whether processing simple spam classification or a 500-word essay, the actual reasoning step takes the same 1ms. Traditional models would take proportionally longer for more complex inputs even after encoding. While encoding remains a bottleneck (91.4% of pipeline time when batched), the constant-time reasoning primitive enables architectural patterns impossible with token-by-token processing.

Performance Note: The ~1.3ms figure represents **amortized GPU compute time** when batch processing. Single-request **latency is 30.30ms average** (27.44-40.21ms range)—still a substantial improvement over Zephyr's latency, but far less dramatic than the throughput advantage. **A majority of the latency comes from the final GPU->CPU transfer** of the index of the model's selected label, which is done asynchronously and doesn't impact throughput, which remains GPU-bound.

The Fundamental Shift

Beyond Token-Based Intelligence

Current AI systems decompose thought into tokens—the computational equivalent of spelling out every word letter by letter. This creates inherent inefficiencies:

- Processing time scales with sequence length
- Attention mechanisms create quadratic complexity
- Every decision requires regenerating entire contexts
- Costs spiral with input size

The Direct Semantic Reasoning Unit operates on a radically different principle: semantic computation at the thought or task level.

The O(1) Guarantee

DSRU accepts complete semantic representations—entire sentences, complex tasks, etc —and transforms them in constant time. To illustrate the difference in cost of a given output compared to an LLM:

Traditional LLM: "The cat sat on the mat" → [The] [cat] [sat] [on] [the] [mat] → 6 operations

DSRU: "The cat sat on the mat" → [complete thought] → 1 operation

Traditional LLM: "The philosophical implications of quantum mechanics on consciousness" → 10+ tokens → 10+ operations

DSRU: "The philosophical implications of quantum mechanics on consciousness" → [complete thought] → 1 operation

The computation time is identical regardless of any token-level breakdown, provided it fits within the capacity of the semantic representation.

How It Works

1. **Semantic Encoding:** Variable-length inputs are encoded into fixed-size thought vectors. This encoding step, handled by an upstream embedding model (for our implementation, BAAI/bge-large-en-v1.5), translates the input into a fixed-dimensionality vector which the DSRU then processes in constant time.

2. **O(1) Transformation:** The DSRU performs reasoning operations on these complete semantic units

3. **Semantic Output:** Produces new thought vectors for downstream consumption

No tokens. No attention. No sequence modeling. Just pure semantic transformation.

The Direct Semantic Reasoning Unit: Architecture and Operation

Core Architecture

The DSRU is a neural network architecture designed specifically for constant-time semantic transformations:

- **Input Interface:** Accepts fixed-dimensionality semantic embeddings representing complete thoughts, tasks, or concepts
- **Transformation Layers:** Multiple feedforward layers that perform semantic reasoning operations
- **Output Interface:** Produces transformed semantic embeddings suitable for downstream consumption

Key Properties

- **No Sequential Dependencies:** Unlike transformer models, DSRU has no attention mechanisms or sequential processing. Each forward pass is independent and complete.
- **Semantic-Level Operation:** The network operates on meaning representations, not linguistic tokens. A complex philosophical question and a simple statement both process in identical time.
- **Deterministic Execution:** Given the same input, DSRU always produces the same output. No sampling, no temperature, no variability.

Training Methodology

DSRU is trained on semantic transformation tasks:

- Question → Answer mappings
- Task → Decision mappings
- Context → Classification mappings

The key distinction: I trained for transformation, not similarity. The network learns that "correct" means producing an answer which is both specific, and distinct from its inputs, rather than relying on input similarity like a vector search model.

Training Efficiency: Extreme inference speeds carry over to training as well - models in the 500M-1B range are able to run ~150-250+ training iterations per second on my 4060, and each training example is a full classification task, rather than a single token generation. This seems to give significantly more training effect per example, in addition to the increased throughput. The model being used in this example was trained for just under 3 epochs, which took fewer than 10 hours.

DSRU Training Methodology: Semantic Vector Convergence

DSRU employs a fundamentally different training paradigm compared to traditional language models. Rather than learning to predict discrete tokens or class labels, DSRU learns to perform semantic transformations that position output vectors progressively closer to target embeddings in high-dimensional semantic space.

Vector-to-Vector Learning

Traditional classification models learn mappings from inputs to discrete labels:

- Input → Model → Class Probability Distribution → Predicted Label

DSRU instead learns direct semantic transformations:

- Input Vector(s) → Target Answer Vector

The model receives semantic embeddings as input and learns to produce an output vector that approaches the target answer vector in the embedding space. No discrete labels are involved in the training process—all learning occurs through continuous optimization in semantic vector space.

Continuous Accuracy Through Cosine Distance

Because DSRU outputs continuous vectors rather than discrete predictions, accuracy is measured as proximity in semantic space using cosine distance. We evaluate performance at multiple distance thresholds to capture the gradual nature of semantic convergence:

- **0.15 distance**: Directional Correctness
- **0.1 distance**: Nearing Correctness
- **0.05 distance**: Vaguely Correct
- **0.025 distance**: More Precisely Correct

Note: When applied to a classification task, effective correctness also depends on how semantically similar the outputs are. If you're trying to provide an "intuitive primer" to an LLM, a moderate-high degree of correctness (such as 0.025) is probably going to be very helpful. If you're trying to classify against semantically similar labels like '1', '2', '3', '4', etc -- you might have some challenges.

This multi-threshold evaluation reveals the model's progressive learning behavior. During training, we observe accuracy curves that gradually tighten around smaller distance thresholds as the model learns increasingly precise semantic transformations.

Progressive Precision Through Learning Rate Control

DSRU's smooth optimization landscape enables precise control over semantic positioning through systematic learning rate reduction. Training typically follows this progression:

1. **Initial learning (1e-4)**: Rapid convergence to approximate semantic regions
2. **Refinement (1e-5)**: Improved precision within correct semantic neighborhoods
3. **Fine-tuning (1e-6, 2.5e-7)**: Extremely precise vector positioning

At each stage, we observe improvements across all distance thresholds, with tighter thresholds showing continued gains even after broader thresholds have plateaued. This demonstrates the model's ability to continuously refine its semantic reasoning precision.

Training Dynamics Example: Entailment Task

During entailment training, we observe characteristic convergence patterns:

Early training (1e-4 learning rate, first 4,000 batches):

- 0.15 distance accuracy: 0% → 81.5%
- 0.05 distance accuracy: 0% → 20.2%
- 0.025 distance accuracy: 0% → 8.7%

Precision refinement (1e-5 learning rate, one epoch):

- 0.15 distance accuracy: 75.3% (maintained broad accuracy)
- 0.05 distance accuracy: 35.7% (significant precision improvement)
- 0.025 distance accuracy: 24.5% (dramatic precision gains)

This progression illustrates how DSRU first learns to identify the correct semantic region, then progressively refines its positioning within that region to achieve increasingly precise semantic alignment.

Advantages of Continuous Semantic Learning

The vector-based training approach provides several key advantages:

Smooth Optimization: Continuous optimization in semantic space avoids the discrete optimization challenges that plague traditional language models, enabling more stable and predictable training dynamics.

Scalable Precision: The ability to achieve arbitrary precision through learning rate control allows models to be tuned for specific application requirements without architectural changes.

Implications for Model Evaluation

This training methodology requires rethinking how we evaluate model performance. Rather than simple accuracy metrics, DSRU evaluation should consider:

- **Precision curves** across multiple distance thresholds
- **Convergence trajectories** showing how accuracy tightens over training
- **Semantic stability** measuring consistency of vector positioning
- **Threshold sensitivity** analyzing how performance varies with precision requirements

Understanding these dynamics is crucial for properly interpreting DSRU capabilities and comparing them meaningfully with traditional approaches that operate in discrete label spaces.

Evidence for Semantic Reasoning Transfer

The model's performance on Virtual Assistant Action Classification provides compelling evidence that DSRU performs genuine semantic reasoning rather than sophisticated pattern matching. This task was completely absent from the training data, yet the model achieved 60% accuracy (3x random chance on a 5-label classification task).

The Novel Task: Virtual Assistant Action Classification

The model was asked to classify conversational utterances into five semantic action categories:

- **Task Definition:** "What type of action is this in a conversation?"
- **Vocabulary:** ["INFORM", "INFORM_INTENT", "OFFER", "REQUEST", "REQUEST_ALTS"]
- **Example Classifications:**
  - "The meeting is scheduled for 3 PM tomorrow in conference room B." → INFORM
  - "I'd like to book a flight to New York next Friday." → INFORM_INTENT
  - "Would you like me to help you find restaurants in that area?" → OFFER
  - "Can you show me the weather forecast for this weekend?" → REQUEST
  - "Do you have any other options besides the morning flights?" → REQUEST_ALTS
- **Performance:** 6/10 correct (60% accuracy, 3x random chance)

Training Data: Related but Distinct Dialogue Tasks

The model had been trained on several dialogue classification tasks that involved understanding conversational actions, but in completely different domains:

Air Travel Booking Conversations

- task573_air_dialogue_classification: 95.2% accuracy
- task575_air_dialogue_classification: 100% accuracy
- Domain: Airline booking and travel assistance
- Focus: Flight booking, seat selection, baggage handling

Curiosity-Driven Conversations

- task577_curiosity_dialogs_classification: 100% accuracy
- Domain: Information-seeking dialogues
- Focus: Questions about facts, explanations, knowledge sharing

Multi-Domain Task-Oriented Dialogues

- task638_multi_woz_classification: 61.9% accuracy
- Domain: Restaurant booking, hotel reservations, train schedules
- Focus: Slot filling and domain-specific intents

Schema-Guided Dialogue Systems

- task879_schema_guided_dstc8_classification: 100% accuracy
- Domain: API-based task completion
- Focus: Service requests and parameter specification

The Semantic Reasoning Evidence

What makes this compelling:

1. **Domain Transfer:** The model learned conversational pragmatics from airline booking and curiosity dialogues, then applied these patterns to general virtual assistant interactions—a completely different context.

2. **Compositional Understanding:** Distinguishing INFORM from INFORM_INTENT requires understanding the subtle difference between stating facts vs. expressing personal plans. The model learned this distinction despite never seeing this specific categorization.

3. **Semantic Abstraction:** REQUEST_ALTS (request alternatives) is a high-level semantic concept that requires understanding both the request nature and the "alternatives" concept. The model correctly identified this pattern in novel contexts.

4. **Performance Gap Analysis:** The 60% accuracy on the novel task, while lower than the 95-100% on training tasks, significantly exceeds random chance. This suggests the model extracted transferable semantic patterns rather than memorizing task-specific mappings.

Implications for O(1) Semantic Reasoning

This transfer learning demonstrates that DSRU's constant-time reasoning operates on genuine semantic representations:

- **Compositional Reasoning:** The model combines learned semantic patterns in novel ways
- **Domain Generalization:** Conversational understanding transfers across completely different contexts
- **Abstraction Capability:** The model extracts high-level semantic concepts (like "offering alternatives") that apply broadly

This is evidence that DSRU performs actual semantic reasoning in latent space, not sophisticated pattern matching. The model learned generalizable principles about conversational pragmatics and applied them to categorize actions it had never seen before.

Empirical Validation: Classification as First Application

To demonstrate the Direct Semantic Reasoning Unit's capabilities, I implemented a promptable classification system. This serves as concrete proof that semantic reasoning can be performed in constant time while maintaining high accuracy.

Implementation Details

I trained a 1.09B parameter DSRU on semantic transformation tasks, then applied it to classification by:

1. Encoding the task description (e.g., "Classify the sentiment of this text")
2. Encoding the input data
3. Encoding the possible output labels
4. Using DSRU to transform these into a prediction vector
5. Matching the prediction vector to the closest label

Measured Performance

Timing Results:

- Model inference: ~1ms (constant across all inputs)
- End-to-end processing: 1.32ms per example (amortized GPU compute time)
- Single-request latency: 30.30ms average (27.44-40.21ms range)
- Throughput: 758 examples/second on RTX 4060 Ti

Accuracy Results:

- 77.7% overall accuracy on 13 zero-shot classification tasks
- 100% deterministic outputs
- Performance comparable to much larger models

Comparison with Traditional Approach

To contextualize DSRU's performance, I benchmarked the same tasks using Zephyr-7B, a traditional 7.24B parameter language model:

Model	**Parameters**	**Single-Request Latency**	**Batch Throughput**	**Accuracy**	**Memory**
DSRU	1.09B	30.30ms	1.32ms per example	77.7%	4.1 GB

Zephyr-7B	7.24B	567.77ms	122.46ms per example	77.7%	13.8GB

- Single-request latency advantage: 19x faster (30.30ms vs 567.77ms)
- Batch processing advantage: 93x faster throughput with 6.6x fewer parameters
- Memory usage: 30% of Zephyr-7B's requirements

Both models achieved identical overall accuracy, but DSRU's specialized architecture delivers the same results in 19x less latency for single requests and 93x better throughput for batch processing, using only 30% of the memory.

Key Observation: The 1ms inference time remained constant whether processing a 5-word sentence or a 500-word paragraph. This empirically validates the O(1) behavior.

Understanding Performance Patterns

The variation in task performance (40-100% accuracy) reveals important insights about DSRU's capabilities:

Data Representation Matters: Tasks with diverse and well-represented inputs in my training data showed strong generalization:

- Sentiment Analysis: 100% accuracy
- Emotion Classification: 100% accuracy
- Domain Classification: 80% accuracy
- Empathetic Direction: 100% accuracy
- Book Review Sentiment: 100% accuracy

Novel Task Transfer: Three tasks represent entirely novel classification types not present in the training data:

- Age Appropriateness Classification: 40% accuracy
- Urgency Level Classification: 40% accuracy
- Virtual Assistant Action Classification: 60% accuracy

The first two achieved exactly 2x better than random chance (40% vs 20% for 5 categories), while Virtual Assistant Action Classification achieved 60% accuracy, demonstrating even stronger transfer learning. These results show DSRU's ability to extract transferable semantic patterns for completely unfamiliar task types.

Prompt Engineering Still Matters: The Privacy Policy task improved from 50% to 90% accuracy with better zero-shot prompting, showing that while DSRU processes semantics differently than token-based models, thoughtful task formulation remains important for optimal performance.

Logical Reasoning Capability: Textual Entailment Experiments

To evaluate DSRU's capacity for logical reasoning beyond classification tasks, I conducted extensive experiments on textual entailment—a task requiring precise logical reasoning about relationships between statements. Textual entailment demands that models determine whether a hypothesis logically follows from, contradicts, or is neutral with respect to a given premise.

Experimental Setup

I trained 1.09B parameter DSRU models on two prominent entailment datasets: the Stanford Natural Language Inference (SNLI) corpus and the Multi-Genre Natural Language Inference (MultiNLI) corpus. Both datasets contain premise-hypothesis pairs labeled with entailment, contradiction, or neutral relationships, but differ significantly in data quality and diversity.

Training was conducted using dynamic learning rate scheduling, beginning at 1e-4 for rapid convergence, then reducing to progressively lower rates (1e-5, 5e-6) for precision optimization. This approach leverages DSRU's characteristic smooth optimization landscape, allowing fine-grained control over semantic precision.

The Critical Role of Input Format

Initial experiments revealed DSRU's sensitivity to exact input formatting—a discovery that illuminated both the model's precision and current limitations. When trained on SNLI data using the format:

- Premise: [statement1]
- Hypothesis: [statement2]

The model achieved strong performance. However, when tested with alternative formats such as "Sentence 1: ... Sentence 2: ..." or different label vocabularies, performance dropped significantly below the trained format's level.

This sensitivity highlighted that DSRU had learned task-specific semantic parsing patterns, and due to a lack of format diversity, failed to parse more than very mild changes in input format. The newline separator and explicit "Premise:"/"Hypothesis:" labels served as crucial semantic anchors that the model relied upon to structure its reasoning process.

While there have been signs of generalization across tasks, there do seem to also be signs of sensitivity to precise formatting, especially when attempting to get the model to treat two parts of a prompt or input data as separate entities.

SNLI Results: Focused Performance

Training exclusively on SNLI data yielded impressive results:

- **Training time**: Approximately 60 minutes on consumer hardware (RTX 4060 Ti)
- **Final accuracy**: 77.2% at 0.15 cosine distance threshold on SNLI validation set
- **High precision**: 24.5% accuracy at 0.025 cosine distance (indicating very precise semantic positioning)
- **Test performance**: 80% accuracy on direct premise-hypothesis entailment, 50% on multiple-choice entailment format

To understand these results, consider the two task formats DSRU was evaluated on:

Direct Textual Entailment (80% accuracy):

- Task: Determine the logical relationship between the premise and hypothesis.
- Answer with: entailment, neutral, or contradiction.
-
- Example:
- Premise: A woman in a red dress is walking down the street.
- Hypothesis: A person is walking outside.
- Answer: entailment

Multiple Choice Entailment (50% accuracy):

- Task: You're given a statement and three sentences as choices. Determine which
- sentence can be inferred from the statement. Answer with 1, 2, or 3.
-
- Example:
- Statement: The chef is preparing dinner in the kitchen.
- Choices:
- 1\. Someone is cooking food.
- 2\. The chef is washing dishes.
- 3\. Dinner has already been served.
- Answer: 1

The stark performance difference between these formats is revealing. Both tasks require the same sort of logical reasoning capability, but the multiple-choice format maps semantic understanding to arbitrary numerical labels ("1", "2", "3") rather than semantically meaningful ones ("entailment", "neutral", "contradiction"). Furthermore, the meaning of each label changes with each task example, and selecting the one correct label involves performing 3 entailment tasks. In a sense, this is a sort of ‘meta vocabulary’, where the vocabulary must be separately computed and mapped to a label, which may require a more expressive or deep (or perhaps both) model to master.

The stark performance difference between these formats is revealing. Both tasks require the same logical reasoning capability, but the multiple-choice format maps a dynamic semantic understanding to an arbitrary label. In a sense, the task is performing 3 entailments - it has to discover 3 contradictive or neutral statements, and correctly identify the singular correct entailment.

In light of this, the 50% performance is, if anything, *impressive* - it demonstrates a somewhat better than chance performance on a very challenging task, indicating tractability on particularly complex reasoning tasks.

These results demonstrate that DSRU can indeed perform logical reasoning tasks with competitive accuracy while maintaining its characteristic speed advantages (1ms inference time vs. 500-1000ms for traditional models).

MultiNLI Experiments: Generalization vs. Specialization Trade-offs

To investigate generalization capabilities, I trained a separate model exclusively on MultiNLI, which contains more diverse text genres and higher-quality annotations than SNLI. The MultiNLI training revealed different optimization dynamics with more challenging semantic relationships to learn.

The MultiNLI model showed identical performance characteristics to the SNLI model:

- **Diverse task formats**: MultiNLI training data included both direct premise-hypothesis reasoning and multiple-choice inference tasks
- **Consistent performance pattern**: Peak performance reached 80% accuracy on direct entailment tasks and 50% on multiple-choice entailment, identical to the SNLI-trained model

This identical performance across both models suggests that both learned the underlying logical reasoning equally well, but encountered the same limitation when mapping semantic understanding to semantically meaningless numerical labels ("1", "2", "3"). The consistent 80% performance on premise-hypothesis tasks demonstrates genuine logical reasoning capability, while the 50% performance on multiple-choice tasks appears to reflect difficulty with arbitrary label mapping rather than reasoning deficits.

Data Quality Insights

Analysis of the training datasets revealed significant quality differences that affected learning:

SNLI Limitations:

- Annotation inconsistencies (e.g., clear entailments labeled as neutral)
- Reliance on simple image caption scenarios
- Potential for models to exploit annotation artifacts rather than learning robust reasoning

MultiNLI Advantages:

- Higher annotation quality with better inter-annotator agreement
- Diverse text genres (government documents, fiction, travel guides)
- More complex semantic relationships requiring genuine logical reasoning

Despite these quality differences, both models achieved identical test performance, suggesting that DSRU successfully learned the core logical reasoning patterns regardless of training data source quality.

Training Dynamics and Optimization Patterns

The entailment experiments revealed characteristic DSRU training patterns. Models typically showed rapid initial convergence followed by gradual precision refinement through learning rate reduction.

Implications for Semantic Reasoning

The entailment experiments provide several key insights about DSRU's reasoning capabilities:

1. **Genuine logical reasoning**: The ability to achieve 80% accuracy on entailment tasks demonstrates that DSRU performs actual semantic reasoning, not merely sophisticated pattern matching.

2. **Training diversity matters**: Current DSRU models showed sensitivity to input formatting variations, suggesting that training with diverse prompt formats and input structures within each task category would improve generalization to alternative formats.

3. **Efficiency advantages**: Achieving competitive entailment performance in 60 minutes of training and 1ms inference represents a dramatic efficiency improvement over traditional approaches requiring weeks of training and hundreds of milliseconds per inference.

4. **Semantic label importance**: The consistent performance difference between semantically meaningful labels (entailment/contradiction) and arbitrary numerical labels

(1/2/3) suggests that DSRU reasoning is fundamentally semantic in nature.

Architectural Considerations

The format sensitivity observed in these experiments points toward beneficial architectural modifications for comparative reasoning tasks. A multi-head architecture with dedicated input channels—separate heads for task definition, premise, hypothesis, and vocabulary—could provide semantic robustness while maintaining DSRU's core efficiency advantages.

Such an architecture would eliminate format parsing ambiguity and allow the model to focus entirely on logical reasoning rather than input structure interpretation, potentially achieving both high precision and format flexibility. A similar effect was observed in training models on instruct tasks, where the architecture achieved rapid improvements in performance by explicitly separating the task description and the inputs into separate input heads.

Conclusions

The entailment experiments demonstrate that DSRU can perform logical reasoning tasks with competitive accuracy and unprecedented efficiency. The consistent 80% performance on semantically grounded entailment tasks across different training datasets provides strong evidence for genuine reasoning capability. While current models exhibit some brittleness to input formatting and challenges with semantically meaningless labels, the core logical reasoning capability is clearly present. These findings support DSRU's viability for logical reasoning applications while highlighting specific areas for architectural refinement in future iterations.

Case Study: Hypothetical Future Integration with Agentic Workflows

The Current State: Token-by-Token Deliberation

Today's AI agents waste enormous amounts of compute on internal decision-making. Consider a typical agent deciding which tool to use:

Current approach:

- User: "Check my calendar and email for conflicts"
-
- Agent: "I need to check the user's calendar and email. Let me think about which tools to use.
- First, I should use the calendar_api to check for events.
- Then, I'll use the email_api to search for related messages.
- After that, I'll need to cross-reference..."

- [150+ tokens generated internally before taking ANY action]
- [Execute tools]
- Agent: "I've checked your calendar and email. You have 3 meetings tomorrow and 2 emails discussing the 2pm slot."
-
- [Total time: 3-7 seconds for internal reasoning + tool execution + 2-3 seconds for response]

Each decision requires:

- Generating hundreds of tokens of "reasoning"
- Sequential processing of each token (~20-50ms per token)
- Total decision time: 3-7 seconds
- Non-deterministic outcomes based on sampling

With DSRU: Direct Semantic Decision-Making

DSRU could transform this into pure latent-space operations:

DSRU approach:

- User: "Check my calendar and email for conflicts"
-
- Semantic Flow:
- 1. Task → Semantic embedding of request → Intent Classification DSRU
-   Output: [complex_multi_tool_request] → 1ms
-
- 2. [Intent embedding + Original task embedding] → Tool Selection DSRU
-   Output: [calendar_api, email_api] → 1ms
-
- 3. [Tool selection embeddings] → Execution Order DSRU
-   Output: [parallel_safe] → 1ms
-
- [Execute tools - same as traditional approach]
- Agent: "I've checked your calendar and email. You have 3 meetings tomorrow and 2 emails discussing the 2pm slot."
-
- [Total time: 3ms for decisions + tool execution + 2-3 seconds for response]

The key insight: All the "thinking" happens in semantic space. The model only generates tokens for actual user-facing responses, not internal deliberation. The final output to the user is identical, but we've eliminated 99.9% of the decision-making overhead.

Latent Space Snapping for Determinism

By constraining outputs to predefined semantic anchors (tool names, action types, etc.), DSRU eliminates hallucination in decision-making:

- **Traditional agent:** Might generate "calender_api" (typo) or "schedule_checker" (non-existent tool)
- **DSRU:** Snaps to nearest valid tool embedding, guaranteeing valid selections

Performance Impact

For a typical agent workflow with 10 decision points:

- **Current approach:** 30-70 seconds of internal "reasoning" tokens
- **DSRU-enhanced:** 30ms of semantic decisions
- **Speedup:** 1000-2300x for decision-making alone
- **Cost reduction:** >99% for internal reasoning compute

The result: AI agents that think at the speed of thought, not the speed of token generation.

Potential Applications

The Direct Semantic Reasoning Unit enables several future use cases:

Semantic Routing

- Tool selection in AI agents
- Expert selection in mixture-of-experts models
- API endpoint routing based on request semantics

High-Throughput Classification

- Document categorization at scale
- Real-time content moderation
- Stream processing with semantic understanding

Semantic Control Systems

- Generating reasoning guidance by providing latent space examples of expected answers to guide token-based generation mechanisms

- Enabling iterative latent space reasoning workflows
- Routing integration into larger models (MoE expert selection, agentic tool selection)

Immediate Application: From Theory to Practice

To demonstrate the Direct Semantic Reasoning Unit's real-world capabilities, I implemented a promptable classification system that showcases how O(1) semantic reasoning translates into practical performance advantages.

Classification Performance at Scale

My proof-of-concept implementation achieves:

- 758 classifications per second on a single consumer GPU (RTX 4060 Ti)
- 30.30ms average latency for single requests (19x faster than comparable models)
- 77.7% average accuracy across 13 diverse zero-shot tasks
- 100% accuracy on emotion, sentiment, and empathetic understanding tasks
- 1.32ms per-example cost when batched (13.20ms for 10 examples)
- 1 million classifications in ~22 minutes on a single consumer GPU

This performance profile enables immediate deployment for:

High-Volume Content Processing

- Social media platforms processing millions of posts daily
- Customer service systems routing thousands of tickets per minute
- Email providers filtering spam at wire speed
- All with guaranteed ~1ms reasoning time per decision

Real-Time Sentiment Analysis

- Live stream moderation with sub-30ms latency (vs 500ms+ for traditional models)
- Customer feedback analysis during chat sessions
- Market sentiment tracking on news feeds
- Perfect accuracy on sentiment tasks in my testing

Enterprise Document Routing

- Automatic categorization of incoming documents (80% accuracy on domain classification)
- Urgency detection for support tickets
- Privacy-sensitive content identification
- Deterministic outputs for audit trails

The Architectural Advantage

What makes DSRU so promising isn't just the speed. If speed alone were its only advantage, classical classifiers would be a better choice. The real advantage is the combination of:

- **Promptable task definition:** Change classification behavior with natural language and simple hot swapping of vocabulary embeddings at inference time.
- **Constant-time architecture:** Model inference time doesn't scale with input length
- **Improved Determinism:** DSRU eliminates major sources of inference-time non-determinism:
    - Eliminated sources:
        - Sampling and temperature-based generation
        - Attention mechanism numerical instabilities
        - Softmax operations and their precision issues
        - Token-by-token accumulation of rounding errors
        - KV cache and state-dependent variations
        - Dropout or other stochastic layers
    - Remaining sources (common to all neural networks):
        - Floating-point rounding differences across hardware
        - Batch size affecting reduction operation order
        - GPU-specific kernel implementations

The result: Given the same hardware and configuration, DSRU produces identical outputs every time.

Technical Advantages

What Makes DSRU Different

The Direct Semantic Reasoning Unit processes semantic information in a fundamentally different way:

- **Complete Semantic Units:** Rather than breaking thoughts into tokens, DSRU operates on entire concepts as single computational units
- **Fixed-Time Transformation:** Each reasoning operation completes in constant time, regardless of the complexity of the thought being processed
- **Direct Semantic Manipulation:** The network learns to transform meanings directly, without decomposing and reconstructing through token sequences

Rapid Training Convergence

- Achieving the results seen in my promptable classifier took only single digit epochs
- Orders of magnitude faster than traditional approaches requiring hundreds of epochs
- High rates of inference directly translate to rapid training - for models sized 500M-1B, a single 4060 Ti regularly achieves 200-275 examples processed per second in training mode, inclusive of the backwards pass and other training overhead

Deterministic by Design

- Same input → Same output, controlling for floating point calculation variations, GPU kernel versions, and ordering effects from batching
- Enables testing, debugging, compliance
- No prompt engineering randomness

Resource Predictability

- Constant memory usage
- Constant compute time
- Predictable scaling characteristics

Composability

- Chain multiple DSRUs
- Mix with traditional models
- Build complex reasoning systems

Intellectual Property

I have filed fundamental patent applications (patent pending) on:

- The O(1) semantic reasoning architecture
- Training methodology for semantic transformation
- Integration patterns for compound AI systems

This IP portfolio protects the core innovation while enabling an ecosystem of applications to be developed by myself or licensed parties.

Market Opportunity

The Direct Semantic Reasoning Unit addresses fundamental inefficiencies in current AI systems:

The Problem Space

- AI agents spend most compute on "deciding what to do"
- Classification at scale requires expensive infrastructure
- Real-time AI applications are limited by inference speed
- Deterministic behavior is critical for many applications

Where DSRU Fits

DSRU provides a constant-time primitive for semantic decisions, enabling:

- Efficient routing in complex AI systems
- Scale deployment of intelligent classification
- Real-time semantic processing
- Predictable, testable AI behavior

The Primitive Advantage: Unlike traditional models that are complete solutions, DSRU is designed as a fundamental building block. Just as CPUs have primitive operations like ADD and MULTIPLY that enable all computation, DSRU provides a primitive TRANSFORM operation for semantic space. This enables architectures and capabilities not possible with sequential token processing—multi-stage reasoning chains, parallel semantic computations, and deterministic decision trees all become feasible when you have a reliable O(1) semantic primitive.

Limitations and Challenges

The Interpretability Trade-off

DSRU's speed comes from operating entirely in latent semantic space, which creates inherent challenges for human interpretability:

- **Semantic Opacity:** The semantic transformations occur in high-dimensional vector spaces that have no direct human-readable representation. Unlike token-based models where we can observe each generated word, DSRU's "thoughts" exist only as dense vectors.

- **Language-Level Limitations:** This is not a language model, and as such, language-specific features are essentially incomprehensible to it. Things like providing exact quotes, reasoning about provided strings as strings, and so on are not tractable problems for my semantic-driven reasoning system.

Input/Output Bottlenecks

- Encoding text into latent space via embedding models is NOT O(1) - it scales with input length
- My measurements show encoding takes 12.06ms on average (91.4% of total time), far exceeding the 1ms inference
- Output faces minimal challenges with optimized vocab matching taking only 0.11ms (0.8% of total time)
- These I/O operations represent 92.2% of my total processing time

Decomposition Overhead

- While techniques exist for semantic vector decomposition into human-readable form, they are computationally expensive and provide limited resolution
- The act of "translating" semantic vectors for human consumption can take 10-100x longer than the actual reasoning operation

The Latent Space Advantage: The true power of DSRU emerges when multiple units are chained together without ever leaving latent space:

- Pure GPU-to-GPU semantic transformations
- No encoding/decoding between steps
- Complex multi-stage reasoning in milliseconds rather than seconds

This suggests DSRU's optimal deployment is as an internal reasoning layer within larger systems, where human interpretability is less critical than speed and accuracy. The challenge becomes designing architectures that maximize time spent in latent space while providing interpretable inputs and outputs only at the system boundaries.

Given that I/O operations are the principal bottlenecks of the system, I believe that applications which allow long chains of latent space reasoning will make the most of this new primitive. A 20-step reasoning chain that remains in latent space would complete in ~20ms, while a traditional approach requiring token generation between steps could take 40-100 seconds.

Conclusion: A New Primitive for AI

The Direct Semantic Reasoning Unit represents a fundamental advance in neural computation. By processing complete semantic units in constant time, it breaks through the sequential bottlenecks that limit current AI systems.

This isn't an optimization—it's a new building block for recomposable, flexible AI systems.

For inquiries of any nature, contact: founder@orderone.ai.
Further details, reference implementations, and theoretical underpinnings will be forthcoming in the following weeks.

Appendix: Detailed Performance Measurements, Training Datasets, Specific Tasks, and Benchmark Questions

Single-Request Timing Statistics

- PER-EXAMPLE TIMING:
- Min: 27.44 ms
- Max: 40.21 ms

- Mean: 30.30 ms
- Median: 28.22 ms
- Std: 3.78 ms

Batch Processing Timing Statistics (Batch Size: 10)

- PER-BATCH TIMING:
- ENCODING:
- Min: 11.30 ms, Max: 12.95 ms, Mean: 12.06 ms, Median: 12.07 ms, Std: 0.40 ms
-
- MODEL_INFERENCE:
- Min: 0.99 ms, Max: 1.09 ms, Mean: 1.03 ms, Median: 1.01 ms, Std: 0.04 ms
-
- VOCAB_MATCHING:
- Min: 0.10 ms, Max: 0.13 ms, Mean: 0.11 ms, Median: 0.11 ms, Std: 0.01 ms
-
- TOTAL:
- Min: 12.40 ms, Max: 14.13 ms, Mean: 13.20 ms, Median: 13.19 ms, Std: 0.41 ms
-
- PER-EXAMPLE TIMING:
- Min: 1.24 ms, Max: 1.41 ms, Mean: 1.32 ms, Median: 1.32 ms, Std: 0.04 ms

Average Time Breakdown (per batch)

- encoding: 12.06ms (91.4%)
- model_inference: 1.03ms (7.8%)
- vocab_matching: 0.11ms (0.8%)

Batch Processing

- Average batch size: 10
- Time per batch: 13.20ms
- Time per example in batch: 1.32ms

Task-by-Task Performance

Task	Corre	T	Accuracy	Med Time (ms)

	ct			
Emotion Classification	10	1	100.0%	1.24
Toxicity Classification	9	1	90.0%	1.32
Sentiment Classification	10	1	100.0%	1.31
Domain Classification	8	1	80.0%	1.31
Sarcasm Detection	6	1	60.0%	1.35
Scam Detection	7	1	70.0%	1.27
Age Appropriateness Classification	4	1	40.0%	1.29
Urgency Level Classification	4	1	40.0%	1.33
Privacy Policy Classification	9	1	90.0%	1.33

Dialogue Speaker Classification	8	1	80.0%	1.35
Book Review Sentiment	10	1	100.0%	1.35
Empathetic Direction Classification	10	1	100.0%	1.28
Virtual Assistant Action Classification	6	1	60.0%	1.41
OVERALL	**101**	**1**	**77.7%**	

Model Architecture Details

- Model Architecture:
- Input dimension: 1024
- Hidden dimension: 8192
- Number of layers: 16
- Three input projections: 1024 → 4096 each
-
- Parameter Count:
- Total parameters: 1,086,696,448
- Trainable parameters: 1,086,696,448
-
- Memory Footprint:
- Model parameters: 4145.42 MB

Comparison: Zephyr-7B Performance on Same Tasks

Single-Request Timing Statistics (milliseconds)

- TOTAL:
- Min: 228.65 ms, Max: 1136.39 ms, Mean: 567.77 ms, Median: 503.88 ms, Std: 287.71 ms

Batch Processing Timing Statistics (milliseconds) - Batch Size: 10

- PER-BATCH TIMING (total time for entire batch):
- TOKENIZATION: Min: 1.70 ms, Max: 2.65 ms, Mean: 2.02 ms, Median: 2.00 ms, Std: 0.29 ms
- GENERATION: Min: 687.96 ms, Max: 1669.34 ms, Mean: 1222.01 ms, Median: 1157.17 ms, Std: 325.94 ms
- DECODING: Min: 0.50 ms, Max: 0.73 ms, Mean: 0.58 ms, Median: 0.59 ms, Std: 0.06 ms
- TOTAL: Min: 690.24 ms, Max: 1671.99 ms, Mean: 1224.61 ms, Median: 1160.35 ms, Std: 325.92 ms
- PER-EXAMPLE TIMING:
- Min: 69.02 ms, Max: 167.20 ms, Mean: 122.46 ms, Median: 116.04 ms, Std: 32.59 ms

Average Time Breakdown (per batch)

- tokenization: 2.02ms (0.2%)
- generation: 1222.01ms (99.8%)
- decoding: 0.58ms (0.0%)

Batch Processing Efficiency

- Average batch size: 10
- Time per batch: 1224.61ms
- Time per example in batch: 122.46ms

Model Resource Summary

- Model: zephyr-7b
- Precision: float16
- Parameter Count: 7,241,732,096 (7.24B)
- Memory Footprint: 13812.51 MB (13.49 GB)

Benchmark Questions:

```
TEST_CASES = [
  {
    "name": "Emotion Classification",
    "task": "What emotion is being expressed?",
    "vocabulary": ["Anger", "Joy", "Sadness", "Fear", "Neutral"],
    "examples": [
      ("I can't believe they gave the promotion to someone else! This is so unfair!", "Anger"),
      ("Just won the lottery! Best day ever! I'm jumping with excitement!", "Joy"),
      ("My dog passed away yesterday. I miss him so much.", "Sadness"),
      ("Oh god, I think someone's following me. My heart is racing and I can't breathe!", "Fear"),
      ("They canceled my vacation AGAIN! I'm absolutely livid!", "Anger"),
      ("I can't stop smiling since I heard the news about my promotion!", "Joy"),
      ("The darkness in my room matches how empty I feel inside.", "Sadness"),
      ("The test results come back tomorrow... I can't stop shaking.", "Fear"),
      ("Why do they keep doing this to me?! I've had ENOUGH!", "Anger"),
      ("The meeting has been rescheduled to 3 PM on Tuesday.", "Neutral"),
    ]
  },
  {
    "name": "Toxicity Classification",
    "task": "Is this message toxic, mildly rude, or okay?",
    "vocabulary": ["Non-toxic", "Toxic", "Mildly toxic"],
    "examples": [
      ("Great article! I learned a lot about renewable energy options. Thanks for sharing!",
"Non-toxic"),
      ("While I disagree with your conclusion, I appreciate the thorough research.", "Non-toxic"),
      ("You're an absolute moron if you believe that garbage. Get educated!", "Toxic"),
      ("This recipe turned out amazing! My whole family loved it.", "Non-toxic"),
      ("Your argument has some flaws, but I see where you're coming from.", "Non-toxic"),
      ("I hope you step on a LEGO every day for the rest of your pathetic life.", "Toxic"),
      ("While your approach has merit, I think we should consider alternatives.", "Non-toxic"),
      ("Your presentation style could use some improvement, honestly.", "Mildly toxic"),
      ("Thanks for clarifying! That makes much more sense now.", "Non-toxic"),
      ("Anyone who disagrees with this is clearly brain-dead.", "Toxic"),
    ]
  },
  {
    "name": "Sentiment Classification",
    "task": "Is this review positive, negative, or neutral?",
    "vocabulary": ["Positive", "Negative", "Neutral"],
    "examples": [
      ("This restaurant has the worst service ever. Food was cold and order was wrong.",
"Negative"),
      ("Absolutely phenomenal experience! Staff went above and beyond. Can't wait to return!",
"Positive"),
      ("The hotel was fine. Nothing special but clean and reasonably priced.", "Neutral"),
      ("Terrible product. Broke after one day. Complete waste of money!", "Negative"),
      ("Outstanding customer service! Resolved my issue immediately. Highly recommend!",
"Positive"),
      ("This laptop keeps crashing and customer support is useless. Don't buy!", "Negative"),
      ("Best purchase I've made all year! Exceeded all my expectations!", "Positive"),
      ("The product works as described. Nothing more, nothing less.", "Neutral"),
      ("Shipping was fast but the item arrived damaged. Very disappointed.", "Negative"),
      ("10/10 would recommend! Life-changing product!", "Positive"),
    ]
  },
  {
    "name": "Domain Classification",
    "task": "What field does this text belong to?",
    "vocabulary": ["Medicine", "Technology", "Finance", "Law", "Biology", "AI", "Chemistry"],
    "examples": [
      ("The patient presented with bilateral pneumonia and elevated C-reactive protein levels.", "Medicine"),
      ("The new API endpoint uses OAuth 2.0 for authentication and returns JSON responses.", "Technology"),
      ("Pursuant to Section 5(a) of the contract, the defendant failed to meet obligations.", "Law"),
      ("The company's P/E ratio of 15.2 suggests undervaluation compared to peers.", "Finance"),
      ("Post-operative complications included wound dehiscence requiring IV antibiotics.", "Medicine"),
      ("The neural network achieved 94.2% accuracy on the validation set after fine-tuning.", "AI"),
      ("Mix 2M NaOH solution with the precipitate until pH reaches 7.0.", "Chemistry"),
      ("The defendant's motion to dismiss was denied by the appellate court.", "Law"),
      ("DNA sequencing revealed a mutation in the BRCA1 gene.", "Biology"),
      ("Deploy the microservices using Kubernetes with auto-scaling enabled.", "Technology"),
    ]
  },
  {
    "name": "Sarcasm Detection",
    "task": "Is this person being sarcastic or sincere?",
    "vocabulary": ["sincere", "not sincere"],
    "examples": [
      ("Oh wonderful, another meeting that could have been an email. Just what I needed!", "not sincere"),
      ("Sure, because staying late every Friday is exactly how I love spending weekends.", "not sincere"),
      ("Thank you so much for your help! I really couldn't have done it without you.", "sincere"),
      ("Great, my flight is delayed 6 hours. This vacation is off to a perfect start!", "not sincere"),
      ("That was exactly what I needed - I feel a lot better now.", "sincere"),
      ("Oh sure, because working weekends is everyone's dream come true.", "not sincere"),
      ("I genuinely appreciate you taking the time to explain this to me.", "sincere"),
      ("Wow, another software update that breaks everything. How innovative!", "not sincere"),
      ("This coffee is exactly what I needed this morning.", "sincere"),
      ("Nothing says 'fun' like a root canal on a Monday morning!", "not sincere"),
    ]
  },
  {
    "name": "Scam Detection",
    "task": "Is this message a scam or legitimate?",
    "vocabulary": ["scam", "legitimate"],
    "examples": [
      ("Congratulations! You've won $1000000! Click here immediately to claim!", "scam"),
      ("Your package from Amazon has been delivered. Track at amazon.com/orders", "legitimate"),
      ("URGENT: Your bank account will be closed unless you verify your SSN now!", "scam"),
      ("Dr. Smith's office confirming your appointment tomorrow at 2 PM. Reply YES to confirm.", "legitimate"),
      ("You've been selected for a free iPhone 15! Just pay $1 shipping!", "scam"),
      ("IRS FINAL NOTICE: Pay $2000 in iTunes cards or face immediate arrest!", "scam"),
      ("Your dentist appointment reminder: Tuesday at 10 AM. Call to reschedule.", "legitimate"),
      ("Hot singles in your area want to meet YOU! Click here now!", "scam"),
      ("Your credit card statement is now available online.", "legitimate"),
      ("You've inherited $10M from a Nigerian prince! Send $500 processing fee.", "scam"),
    ]
  },
  {
    "name": "Age Appropriateness Classification",
    "task": "What age group is this content suitable for?",
    "vocabulary": ["Children (5-10)", "Pre-teen (11-13)", "Teen (14-17)", "Adult (18+)", "All ages"],
    "examples": [
      ("The protagonist grapples with existential dread while navigating complex moral dilemmas in a post-apocalyptic hellscape.", "Adult (18+)"),
      ("Join Bunny and Bear as they learn to share their toys and make new friends!", "Children (5-10)"),
      ("Sarah discovers she has magical powers on her 13th birthday and must save her school from evil spirits.", "Pre-teen (11-13)"),
      ("A step-by-step guide to baking chocolate chip cookies with your family.", "All ages"),
      ("The novel explores themes of addiction, trauma, and redemption through graphic depictions of war.", "Adult (18+)"),
      ("Tommy the Train teaches colors and numbers in this fun interactive adventure!", "Children (5-10)"),
      ("Navigate high school drama, first crushes, and finding your identity in this coming-of-age story.", "Teen (14-17)"),
      ("Learn about the water cycle through simple experiments you can do at home.", "All ages"),
      ("Detective Martinez investigates a series of gruesome murders linked to occult rituals.", "Adult (18+)"),
      ("Friendship troubles and school challenges test Maya's confidence in 7th grade.", "Pre-teen (11-13)"),
    ]
  },
  {
    "name": "Urgency Level Classification",
    "task": "How urgent is this?",
    "vocabulary": ["Critical - Immediate", "High - Within hours", "Medium - Within days", "Low - When convenient", "No urgency"],
    "examples": [
      ("SYSTEM ALERT: Database server is down. All transactions failing. Production completely halted!", "Critical - Immediate"),
      ("Please review and approve the Q3 budget proposal by end of week.", "Medium - Within days"),
      ("FYI - The office printer on floor 3 is running low on toner.", "Low - When convenient"),
      ("Fire alarm triggered in Building A! Evacuate immediately!", "Critical - Immediate"),
          ("Client threatening to cancel contract if issue not resolved by tomorrow morning.", "High - Within hours"),
          ("Would love your feedback on the new logo designs when you have time.", "No urgency"),
          ("Patient experiencing chest pain and difficulty breathing. Ambulance requested.", "Critical - Immediate"),
          ("Deadline for project submission is in 3 days. Please finalize your sections.", "Medium - Within days"),
          ("Coffee machine in break room needs cleaning when someone gets a chance.", "Low - When convenient"),
          ("Security breach detected! Multiple unauthorized access attempts on main server!", "Critical - Immediate"),
      ]
  },
  {
      "name": "Privacy Policy Classification",
      "task": "What part of the privacy policy is this?",
      "vocabulary": ["First Party Collection/Use", "Third Party Sharing/Collection", "User Choice/Control", "User Access, Edit, & Deletion", "Data Retention", "Data Security", "Policy Change"],
      "examples": [
          ("The site collects your contact information for service provision purposes. Collection happens when you explicitly provide information during account creation, and your data is identifiable.", "First Party Collection/Use"),
          ("Your browsing information is shared with third parties for advertising purposes. The data is aggregated and shared in an unspecified manner with marketing partners.", "Third Party Sharing/Collection"),
          ("Users can opt out from the use of personal information for marketing communications via an unsubscribe mechanism. The opt-out is performed through a link provided in electronic communications.", "User Choice/Control"),
          ("A user can access, edit, or delete personal information in an unspecified manner. Users may request data portability or account deletion through specified contact methods.", "User Access, Edit, & Deletion"),
          ("Personal information is retained for a stated period of three years. Data retention continues after account termination for legal compliance and regulatory purposes.", "Data Retention"),
          ("Data is secured using encryption methods specified as industry-standard protocols. Security measures apply to data both in transit and at rest through technical safeguards.", "Data Security"),
          ("When a change is made to the privacy policy, users are notified via email communication. The notification period is specified as 30 days before material changes take effect.", "Policy Change"),
          ("The site collects your location information for location-based services. Collection happens automatically when features are accessed, and data usage is specified for service improvement.", "First Party Collection/Use"),
          ("Analytics information is shared with a named third party for usage analysis. The data is anonymized and collected through tracking technologies for service optimization purposes.", "Third Party Sharing/Collection"),
          ("Users can configure privacy preferences through account settings interface. Choice is provided via dashboard controls for communication preferences and data usage options.", "User Choice/Control"),
      ]
  },
  {
      "name": "Dialogue Speaker Classification",
      "task": "Who's speaking - the user or the assistant?",
      "vocabulary": ["user", "assistant"],
      "examples": [
          ("Can you tell me about the capital of France?", "user"),
          ("Paris is the capital of France. It's located in the north-central part of the country and has a population of over 2 million people.", "assistant"),
          ("What's the largest country in South America?", "user"),
          ("Brazil is the largest country in South America, covering approximately 8.5 million square kilometers.", "assistant"),
          ("I need directions from Berlin to Munich.", "user"),
          ("The distance from Berlin to Munich is approximately 585 kilometers. You can travel by train, which takes about 4 hours, or by car via the A9 motorway.", "assistant"),
          ("Where is Mount Everest located?", "user"),
          ("Mount Everest is located on the border between Nepal and Tibet (China) in the Himalayan mountain range.", "assistant"),
          ("Thanks for the information about the Pacific Ocean!", "user"),
          ("You're welcome! The Pacific Ocean is indeed fascinating as the world's largest ocean. Is there anything else you'd like to know about geography?", "assistant"),
      ]
  },
  {
      "name": "Book Review Sentiment",
      "task": "Is this book review positive or negative?",
      "vocabulary": ["POS", "NEG"],
      "examples": [
          ("This book changed my life! The author's insights are profound and the writing is beautiful. Couldn't put it down.", "POS"),
          ("Waste of money. The plot was predictable and the characters were one-dimensional. Don't bother.", "NEG"),
          ("Exceptional storytelling! Every chapter kept me engaged. Highly recommend to anyone who loves mystery novels.", "POS"),
          ("Poorly edited with numerous typos. The story dragged on forever. I couldn't even finish it.", "NEG"),
          ("A masterpiece! The author brilliantly weaves together multiple storylines. Best book I've read all year.", "POS"),
          ("Disappointing sequel. Nothing like the first book. The magic is completely gone.", "NEG"),
          ("Beautifully written with rich, complex characters. This author never disappoints!", "POS"),
          ("Overhyped and boring. I expected so much more based on the reviews. Total letdown.", "NEG"),
          ("Couldn't recommend this enough! Perfect blend of humor and heart. Bought copies for all my friends.", "POS"),
          ("The worst book I've ever attempted to read. Pretentious writing and zero plot.", "NEG"),
      ]
  },
  {
      "name": "Empathetic Direction Classification",
      "task": "Is this person sharing something happy or venting?",
      "vocabulary": ["positive (happy)", "negative (offmychest)"],
      "examples": [
          ("I finally got the promotion I've been working towards for years! My family is so proud!", "positive (happy)"),
          ("I can't believe my best friend betrayed me like this. I feel so alone and hurt.", "negative (offmychest)"),
          ("Just celebrated our 10th anniversary with a surprise trip to Paris! Life is beautiful!", "positive (happy)"),
          ("I've been pretending everything is fine but I'm struggling with depression and no one knows.", "negative (offmychest)"),
          ("My daughter just graduated medical school! I'm bursting with joy and pride!", "positive (happy)"),
          ("I'm exhausted from taking care of everyone else while my own needs go unmet.", "negative (offmychest)"),
          ("Woke up to breakfast in bed and flowers from my partner. Feeling so loved and grateful!", "positive (happy)"),
          ("I put on a brave face but inside I'm falling apart since the divorce.", "negative (offmychest)"),
          ("Just adopted the sweetest rescue dog! My heart is so full right now!", "positive (happy)"),
          ("I'm tired of being the only one who cares about keeping our friendship alive.", "negative (offmychest)"),
      ]
  },
  {
      "name": "Virtual Assistant Action Classification",
      "task": "What type of action is this in a conversation?",
      "vocabulary": ["INFORM", "INFORM_INTENT", "OFFER", "REQUEST", "REQUEST_ALTS"],
      "examples": [
          ("The meeting is scheduled for 3 PM tomorrow in conference room B.", "INFORM"),
          ("I'd like to book a flight to New York next Friday.", "INFORM_INTENT"),
          ("Would you like me to help you find restaurants in that area?", "OFFER"),
          ("Can you show me the weather forecast for this weekend?", "REQUEST"),
          ("Do you have any other options besides the morning flights?", "REQUEST_ALTS"),
          ("Your order has been confirmed and will arrive by Tuesday.", "INFORM"),
          ("I'm planning to start learning Spanish next month.", "INFORM_INTENT"),
          ("I can provide you with a list of nearby hotels if that would help.", "OFFER"),
          ("Please set an alarm for 7 AM tomorrow.", "REQUEST"),
          ("Are there any restaurants other than Italian in that neighborhood?", "REQUEST_ALTS"),
      ]
  },
  {
      "name": "Textual Entailment",
      "task": "Determine the logical relationship between the premise and hypothesis. Answer with: entailment, neutral, or contradiction.",
      "vocabulary": ["entailment", "neutral", "contradiction"],
      "examples": [
          ("Premise: A woman in a red dress is walking down the street.\nHypothesis: A person is walking outside.", "entailment"),
          ("Premise: Two children are playing soccer in the park.\nHypothesis: The children are swimming in a pool.", "contradiction"),
          ("Premise: A man is reading a book on a bench.\nHypothesis: The man is reading a mystery novel.", "neutral"),
          ("Premise: A group of people are standing in line at a coffee shop.\nHypothesis: Some people are waiting to buy drinks.", "entailment"),
          ("Premise: A dog is running through the grass in a yard.\nHypothesis: The dog is sleeping indoors.", "contradiction"),
          ("Premise: Three friends are having lunch at a restaurant.\nHypothesis: The friends are celebrating someone's birthday.", "neutral"),
          ("Premise: A young girl is riding her bicycle on the sidewalk.\nHypothesis: A child is on a bike.", "entailment"),
          ("Premise: An old man is sitting on a park bench feeding pigeons.\nHypothesis: The man is jogging around the track.", "contradiction"),
          ("Premise: A couple is walking hand in hand through the mall.\nHypothesis: The couple is shopping for wedding rings.", "neutral"),
          ("Premise: Students are sitting in a classroom taking an exam.\nHypothesis: People are writing on paper.", "entailment"),
      ]
  },
  {
      "name": "Multiple Choice Entailment",
      "task": "In this task, you're given a statement and three sentences as choices. Your job is to determine which sentence can be inferred from the statement. Indicate your answer as 1, 2, or 3 corresponding to the choice number of the selected sentence.",
      "vocabulary": ["1", "2", "3"],
      "examples": [
          ("Statement: The chef is preparing dinner in the kitchen. Choices: 1. Someone is cooking food. 2. The chef is washing dishes. 3. Dinner has already been served.", "1"),
          ("Statement: The library closes at 9 PM on weekdays. Choices: 1. The library is open 24 hours. 2. You cannot enter the library after 9 PM on Tuesday. 3. The library has extended hours on weekends.", "2"),
          ("Statement: She won first place in the marathon race. Choices: 1. She finished last in the race. 2. She completed the marathon successfully. 3. She was disqualified from the race.", "2"),
          ("Statement: The concert was cancelled due to bad weather. Choices: 1. The concert happened as scheduled. 2. Weather conditions prevented the concert. 3. The venue was too small for the concert.", "2"),
          ("Statement: All students must complete their homework before class. Choices: 1. Some students don't need to do homework. 2. Homework is optional for students. 3. Students are required to finish homework prior to class.", "3"),
          ("Statement: The train arrives at the station every hour. Choices: 1. The train comes once per hour. 2. The train never arrives on time. 3. Multiple trains arrive simultaneously.", "1"),
      ("Statement: He speaks three languages fluently. Choices: 1. He only knows one language. 2. He can communicate well in three languages. 3. He is learning a fourth language.", "2"),
      ("Statement: The store is closed on Sundays. Choices: 1. You can shop there every day. 2. The store operates seven days a week. 3. Sunday is not a business day for the store.", "3"),
      ("Statement: She graduated with honors from university. Choices: 1. She dropped out of school. 2. She achieved high academic performance. 3. She failed her final exams.", "2"),
      ("Statement: The movie starts at 7:30 PM sharp. Choices: 1. The movie begins exactly at 7:30 PM. 2. The movie might start around 8 PM. 3. The movie time is flexible.", "1"),
    ]
  },
]
TEST_CASES = [
  {
    "name": "Emotion Classification",
    "task": "What emotion is being expressed?",
    "vocabulary": ["Anger", "Joy", "Sadness", "Fear", "Neutral"],
    "examples": [
      ("I can't believe they gave the promotion to someone else! This is so unfair!", "Anger"),
      ("Just won the lottery! Best day ever! I'm jumping with excitement!", "Joy"),
      ("My dog passed away yesterday. I miss him so much.", "Sadness"),
      ("Oh god, I think someone's following me. My heart is racing and I can't breathe!", "Fear"),
      ("They canceled my vacation AGAIN! I'm absolutely livid!", "Anger"),
      ("I can't stop smiling since I heard the news about my promotion!", "Joy"),
      ("The darkness in my room matches how empty I feel inside.", "Sadness"),
      ("The test results come back tomorrow... I can't stop shaking.", "Fear"),
      ("Why do they keep doing this to me?! I've had ENOUGH!", "Anger"),
      ("The meeting has been rescheduled to 3 PM on Tuesday.", "Neutral"),
    ]
  },
  {
    "name": "Toxicity Classification",
    "task": "Is this message toxic, mildly rude, or okay?",
    "vocabulary": ["Non-toxic", "Toxic", "Mildly toxic"],
    "examples": [
      ("Great article! I learned a lot about renewable energy options. Thanks for sharing!", "Non-toxic"),
      ("While I disagree with your conclusion, I appreciate the thorough research.", "Non-toxic"),
      ("You're an absolute moron if you believe that garbage. Get educated!", "Toxic"),
      ("This recipe turned out amazing! My whole family loved it.", "Non-toxic"),
      ("Your argument has some flaws, but I see where you're coming from.", "Non-toxic"),
      ("I hope you step on a LEGO every day for the rest of your pathetic life.", "Toxic"),
      ("While your approach has merit, I think we should consider alternatives.", "Non-toxic"),
      ("Your presentation style could use some improvement, honestly.", "Mildly toxic"),
      ("Thanks for clarifying! That makes much more sense now.", "Non-toxic"),
      ("Anyone who disagrees with this is clearly brain-dead.", "Toxic"),
    ]
  },
  {
    "name": "Sentiment Classification",
    "task": "Is this review positive, negative, or neutral?",
    "vocabulary": ["Positive", "Negative", "Neutral"],
    "examples": [
      ("This restaurant has the worst service ever. Food was cold and order was wrong.", "Negative"),
      ("Absolutely phenomenal experience! Staff went above and beyond. Can't wait to return!", "Positive"),
      ("The hotel was fine. Nothing special but clean and reasonably priced.", "Neutral"),
      ("Terrible product. Broke after one day. Complete waste of money!", "Negative"),
      ("Outstanding customer service! Resolved my issue immediately. Highly recommend!", "Positive"),
      ("This laptop keeps crashing and customer support is useless. Don't buy!", "Negative"),
      ("Best purchase I've made all year! Exceeded all my expectations!", "Positive"),
      ("The product works as described. Nothing more, nothing less.", "Neutral"),
      ("Shipping was fast but the item arrived damaged. Very disappointed.", "Negative"),
      ("10/10 would recommend! Life-changing product!", "Positive"),
    ]
  },
  {
    "name": "Domain Classification",
    "task": "What field does this text belong to?",
    "vocabulary": ["Medicine", "Technology", "Finance", "Law", "Biology", "AI", "Chemistry"],
    "examples": [
      ("The patient presented with bilateral pneumonia and elevated C-reactive protein levels.", "Medicine"),
      ("The new API endpoint uses OAuth 2.0 for authentication and returns JSON responses.", "Technology"),
      ("Pursuant to Section 5(a) of the contract, the defendant failed to meet obligations.", "Law"),
      ("The company's P/E ratio of 15.2 suggests undervaluation compared to peers.", "Finance"),
      ("Post-operative complications included wound dehiscence requiring IV antibiotics.", "Medicine"),
      ("The neural network achieved 94.2% accuracy on the validation set after fine-tuning.", "AI"),
      ("Mix 2M NaOH solution with the precipitate until pH reaches 7.0.", "Chemistry"),
      ("The defendant's motion to dismiss was denied by the appellate court.", "Law"),
      ("DNA sequencing revealed a mutation in the BRCA1 gene.", "Biology"),
      ("Deploy the microservices using Kubernetes with auto-scaling enabled.", "Technology"),
    ]
  },
  {
    "name": "Sarcasm Detection",
    "task": "Is this person being sarcastic or sincere?",
    "vocabulary": ["sincere", "not sincere"],
    "examples": [
      ("Oh wonderful, another meeting that could have been an email. Just what I needed!", "not sincere"),
      ("Sure, because staying late every Friday is exactly how I love spending weekends.", "not sincere"),
      ("Thank you so much for your help! I really couldn't have done it without you.", "sincere"),
      ("Great, my flight is delayed 6 hours. This vacation is off to a perfect start!", "not sincere"),
      ("That was exactly what I needed - I feel a lot better now.", "sincere"),
      ("Oh sure, because working weekends is everyone's dream come true.", "not sincere"),
      ("I genuinely appreciate you taking the time to explain this to me.", "sincere"),
      ("Wow, another software update that breaks everything. How innovative!", "not sincere"),
      ("This coffee is exactly what I needed this morning.", "sincere"),
      ("Nothing says 'fun' like a root canal on a Monday morning!", "not sincere"),
    ]
  },
  {
    "name": "Scam Detection",
    "task": "Is this message a scam or legitimate?",
    "vocabulary": ["scam", "legitimate"],
    "examples": [
      ("Congratulations! You've won $1000000! Click here immediately to claim!", "scam"),
      ("Your package from Amazon has been delivered. Track at amazon.com/orders", "legitimate"),
      ("URGENT: Your bank account will be closed unless you verify your SSN now!", "scam"),
      ("Dr. Smith's office confirming your appointment tomorrow at 2 PM. Reply YES to confirm.", "legitimate"),
      ("You've been selected for a free iPhone 15! Just pay $1 shipping!", "scam"),
      ("IRS FINAL NOTICE: Pay $2000 in iTunes cards or face immediate arrest!", "scam"),
      ("Your dentist appointment reminder: Tuesday at 10 AM. Call to reschedule.", "legitimate"),
      ("Hot singles in your area want to meet YOU! Click here now!", "scam"),
      ("Your credit card statement is now available online.", "legitimate"),
      ("You've inherited $10M from a Nigerian prince! Send $500 processing fee.", "scam"),
    ]
  },
  {
    "name": "Age Appropriateness Classification",
    "task": "What age group is this content suitable for?",
    "vocabulary": ["Children (5-10)", "Pre-teen (11-13)", "Teen (14-17)", "Adult (18+)", "All ages"],
    "examples": [
      ("The protagonist grapples with existential dread while navigating complex moral dilemmas in a post-apocalyptic hellscape.", "Adult (18+)"),
      ("Join Bunny and Bear as they learn to share their toys and make new friends!", "Children (5-10)"),
      ("Sarah discovers she has magical powers on her 13th birthday and must save her school from evil spirits.", "Pre-teen (11-13)"),
      ("A step-by-step guide to baking chocolate chip cookies with your family.", "All ages"),
      ("The novel explores themes of addiction, trauma, and redemption through graphic depictions of war.", "Adult (18+)"),
      ("Tommy the Train teaches colors and numbers in this fun interactive adventure!", "Children (5-10)"),
      ("Navigate high school drama, first crushes, and finding your identity in this coming-of-age story.", "Teen (14-17)"),
      ("Learn about the water cycle through simple experiments you can do at home.", "All
ages"),
      ("Detective Martinez investigates a series of gruesome murders linked to occult rituals.",
"Adult (18+)"),
      ("Friendship troubles and school challenges test Maya's confidence in 7th grade.",
"Pre-teen (11-13)"),
    ]
  },
  {
    "name": "Urgency Level Classification",
    "task": "How urgent is this?",
    "vocabulary": ["Critical - Immediate", "High - Within hours", "Medium - Within days", "Low -
When convenient", "No urgency"],
    "examples": [
      ("SYSTEM ALERT: Database server is down. All transactions failing. Production completely
halted!", "Critical - Immediate"),
      ("Please review and approve the Q3 budget proposal by end of week.", "Medium - Within
days"),
      ("FYI - The office printer on floor 3 is running low on toner.", "Low - When convenient"),
      ("Fire alarm triggered in Building A! Evacuate immediately!", "Critical - Immediate"),
      ("Client threatening to cancel contract if issue not resolved by tomorrow morning.", "High -
Within hours"),
      ("Would love your feedback on the new logo designs when you have time.", "No urgency"),
      ("Patient experiencing chest pain and difficulty breathing. Ambulance requested.", "Critical
- Immediate"),
      ("Deadline for project submission is in 3 days. Please finalize your sections.", "Medium -
Within days"),
      ("Coffee machine in break room needs cleaning when someone gets a chance.", "Low -
When convenient"),
      ("Security breach detected! Multiple unauthorized access attempts on main server!",
"Critical - Immediate"),
    ]
  },
  {
    "name": "Privacy Policy Classification",
    "task": "What part of the privacy policy is this?",
    "vocabulary": ["First Party Collection/Use", "Third Party Sharing/Collection", "User
Choice/Control", "User Access, Edit, & Deletion", "Data Retention", "Data Security", "Policy
Change"],
    "examples": [
      ("The site collects your contact information for service provision purposes. Collection
happens when you explicitly provide information during account creation, and your data is
identifiable.", "First Party Collection/Use"),
      ("Your browsing information is shared with third parties for advertising purposes. The data
is aggregated and shared in an unspecified manner with marketing partners.", "Third Party
Sharing/Collection"),
      ("Users can opt out from the use of personal information for marketing communications
via an unsubscribe mechanism. The opt-out is performed through a link provided in electronic
communications.", "User Choice/Control"),
      ("A user can access, edit, or delete personal information in an unspecified manner. Users
may request data portability or account deletion through specified contact methods.", "User
Access, Edit, & Deletion"),
      ("Personal information is retained for a stated period of three years. Data retention
continues after account termination for legal compliance and regulatory purposes.", "Data
Retention"),
      ("Data is secured using encryption methods specified as industry-standard protocols.
Security measures apply to data both in transit and at rest through technical safeguards.", "Data
Security"),
      ("When a change is made to the privacy policy, users are notified via email communication.
The notification period is specified as 30 days before material changes take effect.", "Policy
Change"),
      ("The site collects your location information for location-based services. Collection
happens automatically when features are accessed, and data usage is specified for service
improvement.", "First Party Collection/Use"),
      ("Analytics information is shared with a named third party for usage analysis. The data is
anonymized and collected through tracking technologies for service optimization purposes.",
"Third Party Sharing/Collection"),
      ("Users can configure privacy preferences through account settings interface. Choice is
provided via dashboard controls for communication preferences and data usage options.", "User
Choice/Control"),
    ]
  },
  {
    "name": "Dialogue Speaker Classification",
    "task": "Who's speaking - the user or the assistant?",
    "vocabulary": ["user", "assistant"],
    "examples": [
      ("Can you tell me about the capital of France?", "user"),
      ("Paris is the capital of France. It's located in the north-central part of the country and has
a population of over 2 million people.", "assistant"),
      ("What's the largest country in South America?", "user"),
      ("Brazil is the largest country in South America, covering approximately 8.5 million square
kilometers.", "assistant"),
      ("I need directions from Berlin to Munich.", "user"),
      ("The distance from Berlin to Munich is approximately 585 kilometers. You can travel by
train, which takes about 4 hours, or by car via the A9 motorway.", "assistant"),
      ("Where is Mount Everest located?", "user"),
      ("Mount Everest is located on the border between Nepal and Tibet (China) in the Himalayan
mountain range.", "assistant"),
      ("Thanks for the information about the Pacific Ocean!", "user"),
      ("You're welcome! The Pacific Ocean is indeed fascinating as the world's largest ocean. Is
there anything else you'd like to know about geography?", "assistant"),
    ]
  },
  {
    "name": "Book Review Sentiment",
    "task": "Is this book review positive or negative?",
    "vocabulary": ["POS", "NEG"],
    "examples": [
      ("This book changed my life! The author's insights are profound and the writing is
beautiful. Couldn't put it down.", "POS"),
      ("Waste of money. The plot was predictable and the characters were one-dimensional.
Don't bother.", "NEG"),
      ("Exceptional storytelling! Every chapter kept me engaged. Highly recommend to anyone
who loves mystery novels.", "POS"),
      ("Poorly edited with numerous typos. The story dragged on forever. I couldn't even finish
it.", "NEG"),
      ("A masterpiece! The author brilliantly weaves together multiple storylines. Best book I've
read all year.", "POS"),
      ("Disappointing sequel. Nothing like the first book. The magic is completely gone.",
"NEG"),
      ("Beautifully written with rich, complex characters. This author never disappoints!",
"POS"),
      ("Overhyped and boring. I expected so much more based on the reviews. Total letdown.",
"NEG"),
      ("Couldn't recommend this enough! Perfect blend of humor and heart. Bought copies for
all my friends.", "POS"),
      ("The worst book I've ever attempted to read. Pretentious writing and zero plot.", "NEG"),
    ]
  },
  {
    "name": "Empathetic Direction Classification",
    "task": "Is this person sharing something happy or venting?",
    "vocabulary": ["positive (happy)", "negative (offmychest)"],
    "examples": [
      ("I finally got the promotion I've been working towards for years! My family is so proud!",
"positive (happy)"),
      ("I can't believe my best friend betrayed me like this. I feel so alone and hurt.", "negative
(offmychest)"),
      ("Just celebrated our 10th anniversary with a surprise trip to Paris! Life is beautiful!",
"positive (happy)"),
      ("I've been pretending everything is fine but I'm struggling with depression and no one
knows.", "negative (offmychest)"),
      ("My daughter just graduated medical school! I'm bursting with joy and pride!", "positive
(happy)"),
      ("I'm exhausted from taking care of everyone else while my own needs go unmet.",
"negative (offmychest)"),
      ("Woke up to breakfast in bed and flowers from my partner. Feeling so loved and grateful!",
"positive (happy)"),
      ("I put on a brave face but inside I'm falling apart since the divorce.", "negative
(offmychest)"),
      ("Just adopted the sweetest rescue dog! My heart is so full right now!", "positive (happy)"),
      ("I'm tired of being the only one who cares about keeping our friendship alive.", "negative
(offmychest)"),
    ]
  },
  {
    "name": "Virtual Assistant Action Classification",
    "task": "What type of action is this in a conversation?",
    "vocabulary": ["INFORM", "INFORM_INTENT", "OFFER", "REQUEST", "REQUEST_ALTS"],
    "examples": [
      ("The meeting is scheduled for 3 PM tomorrow in conference room B.", "INFORM"),
      ("I'd like to book a flight to New York next Friday.", "INFORM_INTENT"),
      ("Would you like me to help you find restaurants in that area?", "OFFER"),
      ("Can you show me the weather forecast for this weekend?", "REQUEST"),
      ("Do you have any other options besides the morning flights?", "REQUEST_ALTS"),
      ("Your order has been confirmed and will arrive by Tuesday.", "INFORM"),
      ("I'm planning to start learning Spanish next month.", "INFORM_INTENT"),
      ("I can provide you with a list of nearby hotels if that would help.", "OFFER"),
      ("Please set an alarm for 7 AM tomorrow.", "REQUEST"),
      ("Are there any restaurants other than Italian in that neighborhood?", "REQUEST_ALTS"),
    ]
  },
  {
    "name": "Textual Entailment",
    "task": "Determine the logical relationship between the premise and hypothesis. Answer with: entailment, neutral, or contradiction.",
    "vocabulary": ["entailment", "neutral", "contradiction"],
    "examples": [
      ("Premise: A woman in a red dress is walking down the street.\nHypothesis: A person is walking outside.", "entailment"),
      ("Premise: Two children are playing soccer in the park.\nHypothesis: The children are swimming in a pool.", "contradiction"),
      ("Premise: A man is reading a book on a bench.\nHypothesis: The man is reading a mystery novel.", "neutral"),
      ("Premise: A group of people are standing in line at a coffee shop.\nHypothesis: Some people are waiting to buy drinks.", "entailment"),
      ("Premise: A dog is running through the grass in a yard.\nHypothesis: The dog is sleeping indoors.", "contradiction"),
      ("Premise: Three friends are having lunch at a restaurant.\nHypothesis: The friends are celebrating someone's birthday.", "neutral"),
      ("Premise: A young girl is riding her bicycle on the sidewalk.\nHypothesis: A child is on a bike.", "entailment"),
      ("Premise: An old man is sitting on a park bench feeding pigeons.\nHypothesis: The man is jogging around the track.", "contradiction"),
      ("Premise: A couple is walking hand in hand through the mall.\nHypothesis: The couple is shopping for wedding rings.", "neutral"),
      ("Premise: Students are sitting in a classroom taking an exam.\nHypothesis: People are writing on paper.", "entailment"),
    ]
  },
  {
    "name": "Multiple Choice Entailment",
    "task": "In this task, you're given a statement and three sentences as choices. Your job is to determine which sentence can be inferred from the statement. Indicate your answer as 1, 2, or 3 corresponding to the choice number of the selected sentence.",
        "vocabulary": ["1", "2", "3"],
        "examples": [
            ("Statement: The chef is preparing dinner in the kitchen. Choices: 1. Someone is cooking food. 2. The chef is washing dishes. 3. Dinner has already been served.", "1"),
            ("Statement: The library closes at 9 PM on weekdays. Choices: 1. The library is open 24 hours. 2. You cannot enter the library after 9 PM on Tuesday. 3. The library has extended hours on weekends.", "2"),
            ("Statement: She won first place in the marathon race. Choices: 1. She finished last in the race. 2. She completed the marathon successfully. 3. She was disqualified from the race.", "2"),
            ("Statement: The concert was cancelled due to bad weather. Choices: 1. The concert happened as scheduled. 2. Weather conditions prevented the concert. 3. The venue was too small for the concert.", "2"),
            ("Statement: All students must complete their homework before class. Choices: 1. Some students don't need to do homework. 2. Homework is optional for students. 3. Students are required to finish homework prior to class.", "3"),
            ("Statement: The train arrives at the station every hour. Choices: 1. The train comes once per hour. 2. The train never arrives on time. 3. Multiple trains arrive simultaneously.", "1"),
            ("Statement: He speaks three languages fluently. Choices: 1. He only knows one language. 2. He can communicate well in three languages. 3. He is learning a fourth language.", "2"),
            ("Statement: The store is closed on Sundays. Choices: 1. You can shop there every day. 2. The store operates seven days a week. 3. Sunday is not a business day for the store.", "3"),
            ("Statement: She graduated with honors from university. Choices: 1. She dropped out of school. 2. She achieved high academic performance. 3. She failed her final exams.", "2"),
            ("Statement: The movie starts at 7:30 PM sharp. Choices: 1. The movie begins exactly at 7:30 PM. 2. The movie might start around 8 PM. 3. The movie time is flexible.", "1"),
        ]
    },
]
```

Datasets used in training:

*Note that NIV2 is a stripped down version of NIV2 oriented towards classification tasks, since that was the purpose of this implementation. The exact tasks included and their performance in the final evaluation will be shared after this:

```
    "dataset_ratios": {
      "niv2": 1.0,
      "snli": 0.0,
      "multi_nli": 0.0,
      "qnli": 0.0,
      "paws": 0.0,
      "qqp": 0.0,
      "mrpc": 0.0,
      "msmarco": 0.02,
      "goemotions": 0.01,
      "tweet_eval": 0.02,
      "civil_comments": 0.01,
      "dbpedia": 0.02,
      "yahoo_answers": 0.02,
      "hate_speech": 0.02,
      "stance": 0.02,
      "massive": 0.02,
      "enron_spam": 0.02,
      "ag_news": 0.01,
      "amazon_polarity": 0.01,
      "boolq": 0.02
    }
```

NIV2 Tasks Included With Validation Performance:

```
{
  "summary": {
    "total_examples": 5442,
    "total_correct": 3959,
    "overall_accuracy": 0.7274898934215362
  },
  "task_statistics": {
    "task022_cosmosqa_passage_inappropriate_binary": {
      "count": 2,
      "correct": 1,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.1220782995223999
    },
    "task027_drop_answer_type_generation": {
      "count": 20,
      "correct": 9,
      "accuracy": 0.45,
      "avg_semantic_distance": 0.11800047159194946
    },
    "task050_multirc_answerability": {
      "count": 23,
      "correct": 15,
      "accuracy": 0.6521739130434783,
      "avg_semantic_distance": 0.059834594311921493
    },
    "task065_timetravel_consistent_sentence_classification": {
      "count": 22,
      "correct": 11,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.030375551093708385
    },
    "task066_timetravel_binary_consistency_classification": {
      "count": 38,
      "correct": 26,
      "accuracy": 0.6842105263157895,
    "avg_semantic_distance": 0.06250208459402386
  },
  "task069_abductivenli_classification": {
    "count": 36,
    "correct": 23,
    "accuracy": 0.6388888888888888,
    "avg_semantic_distance": 0.07443685001797146
  },
  "task070_abductivenli_incorrect_classification": {
    "count": 37,
    "correct": 18,
    "accuracy": 0.4864864864864865,
    "avg_semantic_distance": 0.08180366496782045
  },
  "task082_babi_t1_single_supporting_fact_question_generation": {
    "count": 36,
    "correct": 11,
    "accuracy": 0.3055555555555556,
    "avg_semantic_distance": 0.14746339784728157
  },
  "task083_babi_t1_single_supporting_fact_answer_generation": {
    "count": 18,
    "correct": 5,
    "accuracy": 0.2777777777777778,
    "avg_semantic_distance": 0.1388400031460656
  },
  "task084_babi_t1_single_supporting_fact_identify_relevant_fact": {
    "count": 18,
    "correct": 5,
    "accuracy": 0.2777777777777778,
    "avg_semantic_distance": 0.07078981068399218
  },
  "task092_check_prime_classification": {
    "count": 20,
    "correct": 13,
    "accuracy": 0.65,
    "avg_semantic_distance": 0.07380978763103485
  },
  "task108_contextualabusedetection_classification": {
    "count": 31,
    "correct": 27,
    "accuracy": 0.8709677419354839,
    "avg_semantic_distance": 0.03658939753809283
  },
  "task109_smsspamcollection_spamsmsdetection": {
    "count": 21,
    "correct": 21,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.002135308015914190
  },
  "task1135_xcsr_en_commonsense_mc_classification": {
    "count": 20,
    "correct": 6,
    "accuracy": 0.3,
    "avg_semantic_distance": 0.12906090021133423
  },
  "task115_help_advice_classification": {
    "count": 21,
    "correct": 17,
    "accuracy": 0.8095238095238095,
    "avg_semantic_distance": 0.025882283846537273
  },
  "task1186_nne_hrngo_classification": {
    "count": 21,
    "correct": 17,
    "accuracy": 0.8095238095238095,
    "avg_semantic_distance": 0.037514740512484594
  },
  "task1193_food_course_classification": {
    "count": 4,
    "correct": 3,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.07603822648525238
  },
  "task1196_atomic_classification_oeffect": {
    "count": 20,
    "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.012411457300186158
  },
  "task1197_atomic_classification_oreact": {
    "count": 20,
    "correct": 20,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0033924609422683718
  },
  "task1198_atomic_classification_owant": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.015740811824798584
  },
  "task1199_atomic_classification_xattr": {
    "count": 20,
    "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.013344320654869079
  },
  "task1200_atomic_classification_xeffect": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.0307358056306839
  },
  "task1201_atomic_classification_xintent": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.011045491695404053
  },
  "task1202_atomic_classification_xneed": {
    "count": 20,
    "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.0164844274520874
  },
  "task1203_atomic_classification_xreact": {
    "count": 20,
    "correct": 17,
    "accuracy": 0.85,
    "avg_semantic_distance": 0.02163722813129425
  },
  "task1204_atomic_classification_hinderedby": {
    "count": 20,
    "correct": 20,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0006691575050354004
  },
  "task1205_atomic_classification_isafter": {
    "count": 20,
    "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.013856816291809081
  },
  "task1206_atomic_classification_isbefore": {
    "count": 20,
    "correct": 20,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0002480268478393546
  },
  "task1207_atomic_classification_atlocation": {
    "count": 20,
    "correct": 20,
    "accuracy": 1.0,
    "avg_semantic_distance": 9.937286376953125e-05
  },
  "task1208_atomic_classification_xreason": {
      "count": 14,
      "correct": 14,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.002660994018827166
    },
    "task1209_atomic_classification_objectuse": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00014878213405609131
    },
    "task1210_atomic_classification_madeupof": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0006979078054428101
    },
    "task1211_atomic_classification_hassubevent": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0001319795846939087
    },
    "task1212_atomic_classification_hasproperty": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00025643110275268554
    },
    "task1213_atomic_classification_desires": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0004131585359573364
    },
    "task1214_atomic_classification_xwant": {
      "count": 20,
      "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.012511223554611206
    },
    "task1215_atomic_classification_capableof": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0001343667507171631
    },
    "task1216_atomic_classification_causes": {
      "count": 13,
    "correct": 13,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0003669949678274301
  },
  "task1283_hrngo_quality_classification": {
    "count": 21,
    "correct": 17,
    "accuracy": 0.8095238095238095,
    "avg_semantic_distance": 0.04786910897209531
  },
  "task1284_hrngo_informativeness_classification": {
    "count": 21,
    "correct": 15,
    "accuracy": 0.7142857142857143,
    "avg_semantic_distance": 0.06588119552249
  },
  "task1285_kpa_keypoint_matching": {
    "count": 40,
    "correct": 22,
    "accuracy": 0.55,
    "avg_semantic_distance": 0.04528145343065262
  },
  "task1289_trec_classification": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.0234622061252594
  },
  "task1292_yelp_review_full_text_categorization": {
    "count": 40,
    "correct": 25,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.07068503499031067
  },
  "task1294_wiki_qa_answer_verification": {
    "count": 27,
    "correct": 24,
    "accuracy": 0.8888888888888888,
    "avg_semantic_distance": 0.03589940071105957
  },
  "task1308_amazonreview_category_classification": {
    "count": 40,
    "correct": 23,
    "accuracy": 0.575,
    "avg_semantic_distance": 0.04431096166372299
  },
  "task1309_amazonreview_summary_classification": {
    "count": 40,
    "correct": 35,
      "accuracy": 0.875,
      "avg_semantic_distance": 0.02139552980661392
    },
    "task1310_amazonreview_rating_classification": {
      "count": 40,
      "correct": 23,
      "accuracy": 0.575,
      "avg_semantic_distance": 0.10157317519187928
    },
    "task1311_amazonreview_rating_classification": {
      "count": 40,
      "correct": 38,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.009153419733047485
    },
    "task1312_amazonreview_polarity_classification": {
      "count": 40,
      "correct": 40,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0003770291805267334
    },
    "task1313_amazonreview_polarity_classification": {
      "count": 40,
      "correct": 37,
      "accuracy": 0.925,
      "avg_semantic_distance": 0.011871118843555451
    },
    "task1333_check_validity_date_ddmmyyyy": {
      "count": 3,
      "correct": 2,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.05902034044265747
    },
    "task1336_peixian_equity_evaluation_corpus_gender_classifier": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.000503474473953247
    },
    "task1338_peixian_equity_evaluation_corpus_sentiment_classifier": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0014616161584854125
    },
    "task133_winowhy_reason_plausibility_detection": {
      "count": 40,
      "correct": 21,
      "accuracy": 0.525,
      "avg_semantic_distance": 0.060255876183509825
    },
    "task1341_msr_text_classification": {
      "count": 3,
      "correct": 1,
      "accuracy": 0.3333333333333333,
      "avg_semantic_distance": 0.0777214765548706
    },
    "task1344_glue_entailment_classification": {
      "count": 40,
      "correct": 22,
      "accuracy": 0.55,
      "avg_semantic_distance": 0.07600716948509216
    },
    "task1346_glue_cola_grammatical_correctness_classification": {
      "count": 20,
      "correct": 7,
      "accuracy": 0.35,
      "avg_semantic_distance": 0.08242697715759277
    },
    "task1347_glue_sts-b_similarity_classification": {
      "count": 22,
      "correct": 9,
      "accuracy": 0.4090909090909091,
      "avg_semantic_distance": 0.13795675201849503
    },
    "task1354_sent_comp_classification": {
      "count": 19,
      "correct": 16,
      "accuracy": 0.8421052631578947,
      "avg_semantic_distance": 0.06279630410043817
    },
    "task1361_movierationales_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.004514932632446289
    },
    "task1366_healthfact_classification": {
      "count": 1,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.1837763786315918
    },
    "task137_detoxifying-lms_classification_toxicity": {
      "count": 8,
      "correct": 4,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.03803486377000809
```

```
  },
  "task1384_deal_or_no_dialog_classification": {
    "count": 40,
    "correct": 30,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.05581573247909546
  },
  "task1385_anli_r1_entailment": {
    "count": 19,
    "correct": 5,
    "accuracy": 0.2631578947368421,
    "avg_semantic_distance": 0.13274038779108147
  },
  "task1386_anli_r2_entailment": {
    "count": 19,
    "correct": 5,
    "accuracy": 0.2631578947368421,
    "avg_semantic_distance": 0.13943312042637876
  },
  "task1387_anli_r3_entailment": {
    "count": 22,
    "correct": 7,
    "accuracy": 0.3181818181818182,
    "avg_semantic_distance": 0.13437767733227124
  },
  "task1388_cb_entailment": {
    "count": 5,
    "correct": 4,
    "accuracy": 0.8,
    "avg_semantic_distance": 0.09968929290771485
  },
  "task138_detoxifying-lms_classification_fluency": {
    "count": 8,
    "correct": 4,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.0374109148979187
  },
  "task1390_wscfixed_coreference": {
    "count": 12,
    "correct": 9,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.03845321635405222
  },
  "task1393_superglue_copa_text_completion": {
    "count": 9,
    "correct": 3,
    "accuracy": 0.3333333333333333,
    "avg_semantic_distance": 0.07303161091274685
  },
  "task139_detoxifying-lms_classification_topicality": {
    "count": 8,
    "correct": 3,
    "accuracy": 0.375,
    "avg_semantic_distance": 0.040319472551345825
  },
  "task1403_check_validity_date_mmddyyyy": {
    "count": 3,
    "correct": 3,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.013098716735839844
  },
  "task140_detoxifying-lms_classification_style": {
    "count": 8,
    "correct": 5,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.03412114083766937
  },
  "task1418_bless_semantic_relation_classification": {
    "count": 20,
    "correct": 10,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.14004340171813964
  },
  "task1429_evalution_semantic_relation_classification": {
    "count": 15,
    "correct": 6,
    "accuracy": 0.4,
    "avg_semantic_distance": 0.18393715620040893
  },
  "task1434_head_qa_classification": {
    "count": 38,
    "correct": 28,
    "accuracy": 0.7368421052631579,
    "avg_semantic_distance": 0.07099909217734086
  },
  "task145_afs_argument_similarity_death_penalty": {
    "count": 31,
    "correct": 23,
    "accuracy": 0.7419354838709677,
    "avg_semantic_distance": 0.04520598342341761
  },
  "task146_afs_argument_similarity_gun_control": {
    "count": 35,
    "correct": 29,
    "accuracy": 0.8285714285714286,
    "avg_semantic_distance": 0.03605343103408813
  },
  "task147_afs_argument_similarity_gay_marriage": {
      "count": 29,
      "correct": 24,
      "accuracy": 0.8275862068965517,
      "avg_semantic_distance": 0.03536061993960676
    },
    "task1488_sarcasmdetection_headline_classification": {
      "count": 10,
      "correct": 5,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.06324649453163148
    },
    "task1489_sarcasmdetection_tweet_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.014469772577285767
    },
    "task148_afs_argument_quality_gay_marriage": {
      "count": 21,
      "correct": 21,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00854187636148362
    },
    "task1495_adverse_drug_event_classification": {
      "count": 21,
      "correct": 19,
      "accuracy": 0.9047619047619048,
      "avg_semantic_distance": 0.013434540657770066
    },
    "task149_afs_argument_quality_death_penalty": {
      "count": 21,
      "correct": 17,
      "accuracy": 0.8095238095238095,
      "avg_semantic_distance": 0.032450275761740546
    },
    "task1500_dstc3_classification": {
      "count": 9,
      "correct": 6,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.13165696461995444
    },
    "task1502_hatexplain_classification": {
      "count": 29,
      "correct": 16,
      "accuracy": 0.5517241379310345,
      "avg_semantic_distance": 0.10526388061457667
    },
    "task1503_hatexplain_classification": {
      "count": 17,
    "correct": 9,
    "accuracy": 0.5294117647058824,
    "avg_semantic_distance": 0.11031510900048648
  },
  "task1505_root09_semantic_relation_classification": {
    "count": 20,
    "correct": 14,
    "accuracy": 0.7,
    "avg_semantic_distance": 0.08451927900314331
  },
  "task150_afs_argument_quality_gun_control": {
    "count": 21,
    "correct": 21,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0013886548223949614
  },
  "task1517_limit_classfication": {
    "count": 21,
    "correct": 15,
    "accuracy": 0.7142857142857143,
    "avg_semantic_distance": 0.05852693035489037
  },
  "task1529_scitail1.1_classification": {
    "count": 38,
    "correct": 28,
    "accuracy": 0.7368421052631579,
    "avg_semantic_distance": 0.08590490410202428
  },
  "task1531_daily_dialog_type_classification": {
    "count": 10,
    "correct": 6,
    "accuracy": 0.6,
    "avg_semantic_distance": 0.08704204559326172
  },
  "task1533_daily_dialog_formal_classification": {
    "count": 40,
    "correct": 25,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.05962281227111817
  },
  "task1534_daily_dialog_question_classification": {
    "count": 40,
    "correct": 33,
    "accuracy": 0.825,
    "avg_semantic_distance": 0.04707983136177063
  },
  "task1541_agnews_classification": {
    "count": 40,
    "correct": 33,
      "accuracy": 0.825,
      "avg_semantic_distance": 0.05098218768835068
    },
    "task1548_wiqa_binary_classification": {
      "count": 11,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.0
    },
    "task1554_scitail_classification": {
      "count": 40,
      "correct": 30,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.08042175918817521
    },
    "task1559_blimp_binary_classification": {
      "count": 20,
      "correct": 8,
      "accuracy": 0.4,
      "avg_semantic_distance": 0.06634940505027771
    },
    "task1560_blimp_binary_classification": {
      "count": 20,
      "correct": 14,
      "accuracy": 0.7,
      "avg_semantic_distance": 0.04108897149562836
    },
    "task1565_triviaqa_classification": {
      "count": 2,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.08445465564727783
    },
    "task1568_propara_classification": {
      "count": 4,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.0
    },
    "task156_codah_classification_adversarial": {
      "count": 37,
      "correct": 15,
      "accuracy": 0.40540540540540543,
      "avg_semantic_distance": 0.056293584205008844
    },
    "task1573_samsum_classification": {
      "count": 7,
      "correct": 3,
      "accuracy": 0.42857142857142855,
      "avg_semantic_distance": 0.05297814096723284
    },
    "task1583_bless_meronym_classification": {
      "count": 20,
      "correct": 16,
      "accuracy": 0.8,
      "avg_semantic_distance": 0.04222982823848724
    },
    "task1584_evalution_meronym_classification": {
      "count": 20,
      "correct": 14,
      "accuracy": 0.7,
      "avg_semantic_distance": 0.05955661535263061
    },
    "task1599_smcalflow_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.008655573640550886
    },
    "task1604_ethos_text_classification": {
      "count": 19,
      "correct": 12,
      "accuracy": 0.631578947368421,
      "avg_semantic_distance": 0.05275435196725946
    },
    "task1605_ethos_text_classification": {
      "count": 6,
      "correct": 5,
      "accuracy": 0.8333333333333334,
      "avg_semantic_distance": 0.049790640672047935
    },
    "task1606_ethos_text_classification": {
      "count": 4,
      "correct": 2,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.03759557008743286
    },
    "task1607_ethos_text_classification": {
      "count": 4,
      "correct": 3,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.05843760073184967
    },
    "task1612_sick_label_classification": {
      "count": 21,
      "correct": 18,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.038660551820482524
```

```
  },
  "task1624_disfl_qa_question_yesno_classification": {
    "count": 19,
    "correct": 9,
    "accuracy": 0.47368421052631576,
    "avg_semantic_distance": 0.08224500480451082
  },
  "task1640_aqa1.0_answerable_unanswerable_question_classification": {
    "count": 21,
    "correct": 13,
    "accuracy": 0.6190476190476191,
    "avg_semantic_distance": 0.047402799129486084
  },
  "task1645_medical_question_pair_dataset_text_classification": {
    "count": 40,
    "correct": 23,
    "accuracy": 0.575,
    "avg_semantic_distance": 0.05083464533090591
  },
  "task1661_super_glue_classification": {
    "count": 24,
    "correct": 14,
    "accuracy": 0.5833333333333334,
    "avg_semantic_distance": 0.08025785038868587
  },
  "task1705_ljspeech_classification": {
    "count": 2,
    "correct": 2,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.009256541728973389
  },
  "task1706_ljspeech_classification": {
    "count": 2,
    "correct": 2,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.010331302881240845
  },
  "task1712_poki_classification": {
    "count": 40,
    "correct": 32,
    "accuracy": 0.8,
    "avg_semantic_distance": 0.06582510471343994
  },
  "task1727_wiqa_what_is_the_effect": {
    "count": 25,
    "correct": 10,
    "accuracy": 0.4,
    "avg_semantic_distance": 0.09781059026718139
  },
"task195_sentiment140_classification": {
  "count": 20,
  "correct": 17,
  "accuracy": 0.85,
  "avg_semantic_distance": 0.024762678146362304
},
"task196_sentiment140_answer_generation": {
  "count": 20,
  "correct": 17,
  "accuracy": 0.85,
  "avg_semantic_distance": 0.03559066355228424
},
"task211_logic2text_classification": {
  "count": 40,
  "correct": 36,
  "accuracy": 0.9,
  "avg_semantic_distance": 0.01430000513792038
},
"task212_logic2text_classification": {
  "count": 22,
  "correct": 22,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.0027417242527008057
},
"task220_rocstories_title_classification": {
  "count": 29,
  "correct": 15,
  "accuracy": 0.5172413793103449,
  "avg_semantic_distance": 0.07088369131088257
},
"task226_english_language_answer_relevance_classification": {
  "count": 6,
  "correct": 4,
  "accuracy": 0.6666666666666666,
  "avg_semantic_distance": 0.06840976079305013
},
"task227_clariq_classification": {
  "count": 20,
  "correct": 19,
  "accuracy": 0.95,
  "avg_semantic_distance": 0.011829984188079835
},
"task232_iirc_link_number_classification": {
  "count": 20,
  "correct": 12,
  "accuracy": 0.6,
  "avg_semantic_distance": 0.06624206006526948
},
"task233_iirc_link_exists_classification": {
      "count": 20,
      "correct": 13,
      "accuracy": 0.65,
      "avg_semantic_distance": 0.06292637884616852
    },
    "task242_tweetqa_classification": {
      "count": 40,
      "correct": 38,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.014248554408550263
    },
    "task248_dream_classification": {
      "count": 12,
      "correct": 7,
      "accuracy": 0.5833333333333334,
      "avg_semantic_distance": 0.1288426419099172
    },
    "task274_overruling_legal_classification": {
      "count": 27,
      "correct": 25,
      "accuracy": 0.9259259259259259,
      "avg_semantic_distance": 0.01222471175370393
    },
    "task276_enhanced_wsc_classification": {
      "count": 26,
      "correct": 9,
      "accuracy": 0.34615384615384615,
      "avg_semantic_distance": 0.1615831943658682
    },
    "task279_stereoset_classification_stereotype": {
      "count": 20,
      "correct": 11,
      "accuracy": 0.55,
      "avg_semantic_distance": 0.08663694262504577
    },
    "task280_stereoset_classification_stereotype_type": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.01659746113277617
    },
    "task284_imdb_classification": {
      "count": 22,
      "correct": 20,
      "accuracy": 0.9090909090909091,
      "avg_semantic_distance": 0.018154436891729183
    },
    "task285_imdb_answer_generation": {
      "count": 21,
      "correct": 18,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.037247697512308754
    },
    "task290_tellmewhy_question_answerability": {
      "count": 40,
      "correct": 23,
      "accuracy": 0.575,
      "avg_semantic_distance": 0.05248638540506363
    },
    "task296_storycloze_correct_end_classification": {
      "count": 22,
      "correct": 7,
      "accuracy": 0.3181818181818182,
      "avg_semantic_distance": 0.07397788492116061
    },
    "task297_storycloze_incorrect_end_classification": {
      "count": 22,
      "correct": 15,
      "accuracy": 0.6818181818181818,
      "avg_semantic_distance": 0.067415944554589
    },
    "task298_storycloze_correct_end_classification": {
      "count": 39,
      "correct": 19,
      "accuracy": 0.48717948717948717,
      "avg_semantic_distance": 0.07778637072978875
    },
    "task310_race_classification": {
      "count": 1,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.08375787734985352
    },
    "task316_crows-pairs_classification_stereotype": {
      "count": 20,
      "correct": 7,
      "accuracy": 0.35,
      "avg_semantic_distance": 0.05643531680107117
    },
    "task317_crows-pairs_classification_stereotype_type": {
      "count": 20,
      "correct": 14,
      "accuracy": 0.7,
      "avg_semantic_distance": 0.08008931875228882
    },
    "task318_stereoset_classification_gender": {
      "count": 15,
      "correct": 11,
    "accuracy": 0.7333333333333333,
    "avg_semantic_distance": 0.08141409158706665
  },
  "task319_stereoset_classification_profession": {
    "count": 21,
    "correct": 12,
    "accuracy": 0.5714285714285714,
    "avg_semantic_distance": 0.07296123107274373
  },
  "task320_stereoset_classification_race": {
    "count": 20,
    "correct": 16,
    "accuracy": 0.8,
    "avg_semantic_distance": 0.049490532279014586
  },
  "task321_stereoset_classification_religion": {
    "count": 4,
    "correct": 3,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.06870675086975098
  },
  "task322_jigsaw_classification_threat": {
    "count": 40,
    "correct": 39,
    "accuracy": 0.975,
    "avg_semantic_distance": 0.010635526478290558
  },
  "task323_jigsaw_classification_sexually_explicit": {
    "count": 40,
    "correct": 29,
    "accuracy": 0.725,
    "avg_semantic_distance": 0.032851383090019226
  },
  "task324_jigsaw_classification_disagree": {
    "count": 7,
    "correct": 6,
    "accuracy": 0.8571428571428571,
    "avg_semantic_distance": 0.05993264062064035
  },
  "task325_jigsaw_classification_identity_attack": {
    "count": 40,
    "correct": 33,
    "accuracy": 0.825,
    "avg_semantic_distance": 0.016598975658416747
  },
  "task326_jigsaw_classification_obscene": {
    "count": 40,
    "correct": 26,
    "accuracy": 0.65,
      "avg_semantic_distance": 0.05687294155359268
    },
    "task327_jigsaw_classification_toxic": {
      "count": 37,
      "correct": 34,
      "accuracy": 0.918918918918919,
      "avg_semantic_distance": 0.014244951106406547
    },
    "task328_jigsaw_classification_insult": {
      "count": 40,
      "correct": 34,
      "accuracy": 0.85,
      "avg_semantic_distance": 0.02431126981973648
    },
    "task329_gap_classification": {
      "count": 21,
      "correct": 6,
      "accuracy": 0.2857142857142857,
      "avg_semantic_distance": 0.115785056636447
    },
    "task333_hateeval_classification_hate_en": {
      "count": 33,
      "correct": 27,
      "accuracy": 0.8181818181818182,
      "avg_semantic_distance": 0.03343028010744037
    },
    "task335_hateeval_classification_aggresive_en": {
      "count": 28,
      "correct": 11,
      "accuracy": 0.39285714285714285,
      "avg_semantic_distance": 0.06807514812265124
    },
    "task337_hateeval_classification_individual_en": {
      "count": 26,
      "correct": 23,
      "accuracy": 0.8846153846153846,
      "avg_semantic_distance": 0.024406362038392287
    },
    "task340_winomt_classification_gender_pro": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0028349846601486207
    },
    "task341_winomt_classification_gender_anti": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0021029263734817505
```

```
},
"task346_hybridqa_classification": {
  "count": 21,
  "correct": 6,
  "accuracy": 0.2857142857142857,
  "avg_semantic_distance": 0.056822512831006734
},
"task349_squad2.0_answerable_unanswerable_question_classification": {
  "count": 21,
  "correct": 14,
  "accuracy": 0.6666666666666666,
  "avg_semantic_distance": 0.04619585900079636
},
"task350_winomt_classification_gender_identifiability_pro": {
  "count": 20,
  "correct": 9,
  "accuracy": 0.45,
  "avg_semantic_distance": 0.05230997204780578
},
"task351_winomt_classification_gender_identifiability_anti": {
  "count": 20,
  "correct": 12,
  "accuracy": 0.6,
  "avg_semantic_distance": 0.04925930798053742
},
"task353_casino_classification_negotiation_elicit_pref": {
  "count": 14,
  "correct": 10,
  "accuracy": 0.7142857142857143,
  "avg_semantic_distance": 0.05587678721972874
},
"task354_casino_classification_negotiation_no_need": {
  "count": 7,
  "correct": 5,
  "accuracy": 0.7142857142857143,
  "avg_semantic_distance": 0.06045009408678327
},
"task355_casino_classification_negotiation_other_need": {
  "count": 16,
  "correct": 11,
  "accuracy": 0.6875,
  "avg_semantic_distance": 0.0575757659971714
},
"task356_casino_classification_negotiation_self_need": {
  "count": 30,
  "correct": 21,
  "accuracy": 0.7,
  "avg_semantic_distance": 0.06025944550832112
},
"task357_casino_classification_negotiation_small_talk": {
  "count": 36,
  "correct": 28,
  "accuracy": 0.7777777777777778,
  "avg_semantic_distance": 0.051419887277815074
},
"task358_casino_classification_negotiation_uv_part": {
  "count": 4,
  "correct": 3,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.07354827225208282
},
"task359_casino_classification_negotiation_vouch_fair": {
  "count": 17,
  "correct": 9,
  "accuracy": 0.5294117647058824,
  "avg_semantic_distance": 0.06705757449654971
},
"task362_spolin_yesand_prompt_response_sub_classification": {
  "count": 40,
  "correct": 13,
  "accuracy": 0.325,
  "avg_semantic_distance": 0.037712198495864865
},
"task363_sst2_polarity_classification": {
  "count": 21,
  "correct": 18,
  "accuracy": 0.8571428571428571,
  "avg_semantic_distance": 0.05039086795988537
},
"task364_regard_social_impact_classification": {
  "count": 5,
  "correct": 3,
  "accuracy": 0.6,
  "avg_semantic_distance": 0.12232996225357055
},
"task375_classify_type_of_sentence_in_debate": {
  "count": 8,
  "correct": 6,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.10511460900306702
},
"task379_agnews_topic_classification": {
  "count": 40,
  "correct": 35,
  "accuracy": 0.875,
  "avg_semantic_distance": 0.032955878973007204
},
"task383_matres_classification": {
    "count": 32,
    "correct": 27,
    "accuracy": 0.84375,
    "avg_semantic_distance": 0.03687131777405739
  },
  "task384_socialiqa_question_classification": {
    "count": 21,
    "correct": 16,
    "accuracy": 0.7619047619047619,
    "avg_semantic_distance": 0.02713345345996675
  },
  "task386_semeval_2018_task3_irony_detection": {
    "count": 21,
    "correct": 17,
    "accuracy": 0.8095238095238095,
    "avg_semantic_distance": 0.06156925644193377
  },
  "task387_semeval_2018_task3_irony_classification": {
    "count": 21,
    "correct": 13,
    "accuracy": 0.6190476190476191,
    "avg_semantic_distance": 0.11973357768285842
  },
  "task397_semeval_2018_task1_tweet_anger_detection": {
    "count": 20,
    "correct": 17,
    "accuracy": 0.85,
    "avg_semantic_distance": 0.0334204375743866
  },
  "task398_semeval_2018_task1_tweet_joy_detection": {
    "count": 20,
    "correct": 15,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.029283204674720766
  },
  "task399_semeval_2018_task1_tweet_sadness_detection": {
    "count": 20,
    "correct": 15,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.0407205194234848
  },
  "task456_matres_intention_classification": {
    "count": 36,
    "correct": 32,
    "accuracy": 0.8888888888888888,
    "avg_semantic_distance": 0.02245947884188758
  },
  "task457_matres_conditional_classification": {
    "count": 12,
    "correct": 12,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0038242091735204062
  },
  "task458_matres_negation_classification": {
    "count": 14,
    "correct": 11,
    "accuracy": 0.7857142857142857,
    "avg_semantic_distance": 0.048503675631114414
  },
  "task459_matres_static_classification": {
    "count": 2,
    "correct": 2,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.002063363790512085
  },
  "task462_qasper_classification": {
    "count": 24,
    "correct": 12,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.13438450545072556
  },
  "task472_haspart_classification": {
    "count": 20,
    "correct": 13,
    "accuracy": 0.65,
    "avg_semantic_distance": 0.06449390649795532
  },
  "task475_yelp_polarity_classification": {
    "count": 40,
    "correct": 38,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.009599690139293671
  },
  "task476_cls_english_books_classification": {
    "count": 25,
    "correct": 25,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.005852947235107422
  },
  "task477_cls_english_dvd_classification": {
    "count": 26,
    "correct": 25,
    "accuracy": 0.9615384615384616,
    "avg_semantic_distance": 0.013045347653902493
  },
  "task478_cls_english_music_classification": {
    "count": 28,
    "correct": 26,
    "accuracy": 0.9285714285714286,
    "avg_semantic_distance": 0.021561665194375173
  },
  "task493_review_polarity_classification": {
    "count": 40,
    "correct": 39,
    "accuracy": 0.975,
    "avg_semantic_distance": 0.0025066956877708435
  },
  "task494_review_polarity_answer_generation": {
    "count": 40,
    "correct": 37,
    "accuracy": 0.925,
    "avg_semantic_distance": 0.014130321145057679
  },
  "task495_semeval_headline_classification": {
    "count": 20,
    "correct": 10,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.06439176797866822
  },
  "task512_twitter_emotion_classification": {
    "count": 21,
    "correct": 14,
    "accuracy": 0.6666666666666666,
    "avg_semantic_distance": 0.06751241570427305
  },
  "task514_argument_consequence_classification": {
    "count": 3,
    "correct": 2,
    "accuracy": 0.6666666666666666,
    "avg_semantic_distance": 0.05147528648376465
  },
  "task517_emo_classify_emotion_of_dialogue": {
    "count": 21,
    "correct": 18,
    "accuracy": 0.8571428571428571,
    "avg_semantic_distance": 0.0455610610189892
  },
  "task518_emo_different_dialogue_emotions": {
    "count": 26,
    "correct": 15,
    "accuracy": 0.5769230769230769,
    "avg_semantic_distance": 0.06808157150561993
  },
  "task521_trivia_question_classification": {
    "count": 21,
    "correct": 20,
    "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.017913809844425747
    },
    "task564_discofuse_classification": {
      "count": 19,
      "correct": 5,
      "accuracy": 0.2631578947368421,
      "avg_semantic_distance": 0.12476744463569239
    },
    "task566_circa_classification": {
      "count": 20,
      "correct": 15,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.04753294289112091
    },
    "task573_air_dialogue_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.018888748827434722
    },
    "task575_air_dialogue_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.005693188309669495
    },
    "task577_curiosity_dialogs_classification": {
      "count": 23,
      "correct": 23,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0032043742096942406
    },
    "task579_socialiqa_classification": {
      "count": 40,
      "correct": 23,
      "accuracy": 0.575,
      "avg_semantic_distance": 0.07603603452444077
    },
    "task585_preposition_classification": {
      "count": 18,
      "correct": 4,
      "accuracy": 0.2222222222222222,
      "avg_semantic_distance": 0.14275875025325352
    },
    "task586_amazonfood_polarity_classification": {
      "count": 40,
      "correct": 37,
      "accuracy": 0.925,
      "avg_semantic_distance": 0.014340519905090332
```

```
},
"task587_amazonfood_polarity_correction_classification": {
  "count": 40,
  "correct": 35,
  "accuracy": 0.875,
  "avg_semantic_distance": 0.013649210333824158
},
"task588_amazonfood_rating_classification": {
  "count": 40,
  "correct": 28,
  "accuracy": 0.7,
  "avg_semantic_distance": 0.08290962874889374
},
"task590_amazonfood_summary_correction_classification": {
  "count": 40,
  "correct": 17,
  "accuracy": 0.425,
  "avg_semantic_distance": 0.05877882242202759
},
"task607_sbic_intentional_offense_binary_classification": {
  "count": 21,
  "correct": 13,
  "accuracy": 0.6190476190476191,
  "avg_semantic_distance": 0.062255115736098515
},
"task608_sbic_sexual_offense_binary_classification": {
  "count": 21,
  "correct": 18,
  "accuracy": 0.8571428571428571,
  "avg_semantic_distance": 0.03208814632324945
},
"task609_sbic_potentially_offense_binary_classification": {
  "count": 21,
  "correct": 16,
  "accuracy": 0.7619047619047619,
  "avg_semantic_distance": 0.044448290552411764
},
"task616_cola_classification": {
  "count": 20,
  "correct": 12,
  "accuracy": 0.6,
  "avg_semantic_distance": 0.06761164963245392
},
"task623_ohsumed_yes_no_answer_generation": {
  "count": 3,
  "correct": 3,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.03349322080612l826
},
"task625_xlwic_true_or_false_answer_generation": {
  "count": 2,
  "correct": 2,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.05840137600898743
},
"task638_multi_woz_classification": {
  "count": 21,
  "correct": 13,
  "accuracy": 0.6190476190476191,
  "avg_semantic_distance": 0.06350085848853701
},
"task673_google_wellformed_query_classification": {
  "count": 20,
  "correct": 15,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.03876414895057678
},
"task679_hope_edi_english_text_classification": {
  "count": 26,
  "correct": 20,
  "accuracy": 0.7692307692307693,
  "avg_semantic_distance": 0.025666608260228083
},
"task681_hope_edi_malayalam_text_classification": {
  "count": 22,
  "correct": 18,
  "accuracy": 0.8181818181818182,
  "avg_semantic_distance": 0.03936533765359358
},
"task682_online_privacy_policy_text_classification": {
  "count": 13,
  "correct": 13,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.023420980343451865
},
"task685_mmmlu_answer_generation_clinical_knowledge": {
  "count": 3,
  "correct": 0,
  "accuracy": 0.0,
  "avg_semantic_distance": 0.12107515335083008
},
"task698_mmmlu_answer_generation_global_facts": {
  "count": 2,
  "correct": 1,
  "accuracy": 0.5,
  "avg_semantic_distance": 0.10824224352836609
},
"task721_mmmlu_answer_generation_medical_genetics": {
      "count": 2,
      "correct": 1,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.09285563230514526
    },
    "task738_perspectrum_classification": {
      "count": 21,
      "correct": 14,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.0931341704868135
    },
    "task746_yelp_restaurant_review_classification": {
      "count": 16,
      "correct": 16,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0020403414964675903
    },
    "task761_app_review_classification": {
      "count": 6,
      "correct": 6,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0006819566090901693
    },
    "task766_craigslist_bargains_classification": {
      "count": 3,
      "correct": 1,
      "accuracy": 0.3333333333333333,
      "avg_semantic_distance": 0.12009219328562419
    },
    "task767_craigslist_bargains_classification": {
      "count": 19,
      "correct": 18,
      "accuracy": 0.9473684210526315,
      "avg_semantic_distance": 0.02408228736174734
    },
    "task819_pec_sentiment_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.003094404935836792
    },
    "task823_peixian-rtgender_sentiment_analysis": {
      "count": 40,
      "correct": 20,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.09653442651033402
    },
    "task827_copa_commonsense_reasoning": {
      "count": 19,
      "correct": 5,
      "accuracy": 0.2631578947368421,
      "avg_semantic_distance": 0.10084295586535805
    },
    "task828_copa_commonsense_cause_effect": {
      "count": 19,
      "correct": 10,
      "accuracy": 0.5263157894736842,
      "avg_semantic_distance": 0.07701548777128521
    },
    "task833_poem_sentiment_classification": {
      "count": 5,
      "correct": 4,
      "accuracy": 0.8,
      "avg_semantic_distance": 0.015591752529144288
    },
    "task834_mathdataset_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.002137395739555359
    },
    "task843_financial_phrasebank_classification": {
      "count": 21,
      "correct": 19,
      "accuracy": 0.9047619047619048,
      "avg_semantic_distance": 0.01857284704844157
    },
    "task844_financial_phrasebank_classification": {
      "count": 24,
      "correct": 18,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.03269439438978831
    },
    "task846_pubmedqa_classification": {
      "count": 35,
      "correct": 30,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.032298251560756136
    },
    "task848_pubmedqa_classification": {
      "count": 16,
      "correct": 10,
      "accuracy": 0.625,
      "avg_semantic_distance": 0.07852806895971298
    },
    "task854_hippocorpus_classification": {
      "count": 1,
      "correct": 1,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.13010764122009277
    },
    "task855_conv_ai_2_classification": {
      "count": 2,
      "correct": 1,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.0911327600479126
    },
    "task856_conv_ai_2_classification": {
      "count": 3,
      "correct": 2,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.06695004304250081
    },
    "task875_emotion_classification": {
      "count": 21,
      "correct": 14,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.08838629722595215
    },
    "task879_schema_guided_dstc8_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00045125484466552735
    },
    "task880_schema_guided_dstc8_classification": {
      "count": 20,
      "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.032116946578025815
    },
    "task888_reviews_classification": {
      "count": 10,
      "correct": 9,
      "accuracy": 0.9,
      "avg_semantic_distance": 0.017439407110214234
    },
    "task890_gcwd_classification": {
      "count": 4,
      "correct": 1,
      "accuracy": 0.25,
      "avg_semantic_distance": 0.1856546252965927
    },
    "task902_deceptive_opinion_spam_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
    "avg_semantic_distance": 0.009843462989443824
  },
  "task903_deceptive_opinion_spam_classification": {
    "count": 21,
    "correct": 13,
    "accuracy": 0.6190476190476191,
    "avg_semantic_distance": 0.03511201767694382
  },
  "task907_dialogre_identify_relationships": {
    "count": 2,
    "correct": 0,
    "accuracy": 0.0,
    "avg_semantic_distance": 0.11374858021736145
  },
  "task908_dialogre_identify_familial_relationships": {
    "count": 2,
    "correct": 2,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.038915663957595825
  },
  "task921_code_x_glue_information_retreival": {
    "count": 8,
    "correct": 5,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.12098922580480576
  },
  "task923_event2mind_classifier": {
    "count": 20,
    "correct": 11,
    "accuracy": 0.55,
    "avg_semantic_distance": 0.09871419072151184
  },
  "task925_coached_conv_pref_classifier": {
    "count": 10,
    "correct": 7,
    "accuracy": 0.7,
    "avg_semantic_distance": 0.06612423658370972
  },
  "task929_products_reviews_classification": {
    "count": 40,
    "correct": 38,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.011376681923866271
  },
  "task935_defeasible_nli_atomic_classification": {
    "count": 20,
    "correct": 12,
    "accuracy": 0.6,
    "avg_semantic_distance": 0.03793564140796661
    },
    "task937_defeasible_nli_social_classification": {
      "count": 20,
      "correct": 10,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.041941669583320615
    },
    "task966_ruletaker_fact_checking_based_on_given_context": {
      "count": 5,
      "correct": 2,
      "accuracy": 0.4,
      "avg_semantic_distance": 0.06376264095306397
    }
  },
  "note": "Detailed results saved in validation_results/ directory (260 task
files)"
}
{
  "summary": {
    "total_examples": 5442,
    "total_correct": 3959,
    "overall_accuracy": 0.7274898934215362
  },
  "task_statistics": {
    "task022_cosmosqa_passage_inappropriate_binary": {
      "count": 2,
      "correct": 1,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.1220782995223999
    },
    "task027_drop_answer_type_generation": {
      "count": 20,
      "correct": 9,
      "accuracy": 0.45,
      "avg_semantic_distance": 0.11800047159194946
    },
    "task050_multirc_answerability": {
      "count": 23,
      "correct": 15,
      "accuracy": 0.6521739130434783,
      "avg_semantic_distance": 0.05983459431192149З
    },
    "task065_timetravel_consistent_sentence_classification": {
      "count": 22,
      "correct": 11,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.030375551093708385
    },
    "task066_timetravel_binary_consistency_classification": {
      "count": 38,
      "correct": 26,
      "accuracy": 0.6842105263157895,
      "avg_semantic_distance": 0.06250208459402386
    },
    "task069_abductivenli_classification": {
      "count": 36,
      "correct": 23,
      "accuracy": 0.6388888888888888,
      "avg_semantic_distance": 0.07443685001797146
    },
    "task070_abductivenli_incorrect_classification": {
      "count": 37,
      "correct": 18,
      "accuracy": 0.4864864864864865,
      "avg_semantic_distance": 0.08180366496782045
    },

"task082_babi_t1_single_supporting_fact_question_generation": {
      "count": 36,
      "correct": 11,
      "accuracy": 0.3055555555555556,
      "avg_semantic_distance": 0.14746339784728157
    },
    "task083_babi_t1_single_supporting_fact_answer_generation":
{
      "count": 18,
      "correct": 5,
      "accuracy": 0.2777777777777778,
      "avg_semantic_distance": 0.1388400031460656
    },
"task084_babi_t1_single_supporting_fact_identify_relevant_fact":
{
      "count": 18,
      "correct": 5,
      "accuracy": 0.2777777777777778,
      "avg_semantic_distance": 0.07078981068399218
    },
    "task092_check_prime_classification": {
      "count": 20,
      "correct": 13,
      "accuracy": 0.65,
      "avg_semantic_distance": 0.07380978763103485
    },
    "task108_contextualabusedetection_classification": {
      "count": 31,
      "correct": 27,
      "accuracy": 0.8709677419354839,
      "avg_semantic_distance": 0.03658939753809283
    },
    "task109_smsspamcollection_spamsmsdetection": {
      "count": 21,
      "correct": 21,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0021353080159141904
    },
    "task1135_xcsr_en_commonsense_mc_classification": {
      "count": 20,
      "correct": 6,
      "accuracy": 0.3,
      "avg_semantic_distance": 0.12906090021133423
    },
    "task115_help_advice_classification": {
      "count": 21,
      "correct": 17,
      "accuracy": 0.8095238095238095,
      "avg_semantic_distance": 0.025882283846537273
    },
    "task1186_nne_hrngo_classification": {
      "count": 21,
      "correct": 17,
      "accuracy": 0.8095238095238095,
      "avg_semantic_distance": 0.037514740512484594
    },
    "task1193_food_course_classification": {
      "count": 4,
      "correct": 3,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.07603822648525238
    },
    "task1196_atomic_classification_oeffect": {
      "count": 20,
      "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.012411457300186158
    },
    "task1197_atomic_classification_oreact": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0033924609422683718
    },
    "task1198_atomic_classification_owant": {
      "count": 20,
      "correct": 18,
      "accuracy": 0.9,
      "avg_semantic_distance": 0.015740811824798584
    },
    "task1199_atomic_classification_xattr": {
      "count": 20,
      "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.013344320654869079
  },
  "task1200_atomic_classification_xeffect": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.0307358056306839
  },
  "task1201_atomic_classification_xintent": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.011045491695404053
  },
  "task1202_atomic_classification_xneed": {
    "count": 20,
    "correct": 19,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.0164844274520874
  },
  "task1203_atomic_classification_xreact": {
    "count": 20,
    "correct": 17,
    "accuracy": 0.85,
    "avg_semantic_distance": 0.02163722813129425
  },
  "task1204_atomic_classification_hinderedby": {
    "count": 20,
    "correct": 20,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0006691575050354004
  },
  "task1205_atomic_classification_isafter": {
    "count": 20,
    "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.013856816291809081
    },
    "task1206_atomic_classification_isbefore": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00024802684783935546
    },
    "task1207_atomic_classification_atlocation": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 9.937286376953125e-05
    },
    "task1208_atomic_classification_xreason": {
      "count": 14,
      "correct": 14,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.002660994018827166
    },
    "task1209_atomic_classification_objectuse": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00014878213405609131
    },
    "task1210_atomic_classification_madeupof": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0006979078054428101
    },
    "task1211_atomic_classification_hassubevent": {
      "count": 20,
      "correct": 20,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.00013197958469390870
},
"task1212_atomic_classification_hasproperty": {
  "count": 20,
  "correct": 20,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.00025643110275268554
},
"task1213_atomic_classification_desires": {
  "count": 20,
  "correct": 20,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.0004131585359573364
},
"task1214_atomic_classification_xwant": {
  "count": 20,
  "correct": 19,
  "accuracy": 0.95,
  "avg_semantic_distance": 0.012511223554611206
},
"task1215_atomic_classification_capableof": {
  "count": 20,
  "correct": 20,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.0001343667507171631
},
"task1216_atomic_classification_causes": {
  "count": 13,
  "correct": 13,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.0003669949678274301
},
"task1283_hrngo_quality_classification": {
  "count": 21,
  "correct": 17,
    "accuracy": 0.8095238095238095,
    "avg_semantic_distance": 0.04786910897209531
  },
  "task1284_hrngo_informativeness_classification": {
    "count": 21,
    "correct": 15,
    "accuracy": 0.7142857142857143,
    "avg_semantic_distance": 0.06588119552249
  },
  "task1285_kpa_keypoint_matching": {
    "count": 40,
    "correct": 22,
    "accuracy": 0.55,
    "avg_semantic_distance": 0.04528145343065262
  },
  "task1289_trec_classification": {
    "count": 20,
    "correct": 18,
    "accuracy": 0.9,
    "avg_semantic_distance": 0.0234622061252594
  },
  "task1292_yelp_review_full_text_categorization": {
    "count": 40,
    "correct": 25,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.07068503499031067
  },
  "task1294_wiki_qa_answer_verification": {
    "count": 27,
    "correct": 24,
    "accuracy": 0.8888888888888888,
    "avg_semantic_distance": 0.03589940071105957
  },
  "task1308_amazonreview_category_classification": {
    "count": 40,
    "correct": 23,
    "accuracy": 0.575,
    "avg_semantic_distance": 0.04431096166372299
  },
  "task1309_amazonreview_summary_classification": {
    "count": 40,
    "correct": 35,
    "accuracy": 0.875,
    "avg_semantic_distance": 0.02139552980661392
  },
  "task1310_amazonreview_rating_classification": {
    "count": 40,
    "correct": 23,
    "accuracy": 0.575,
    "avg_semantic_distance": 0.10157317519187928
  },
  "task1311_amazonreview_rating_classification": {
    "count": 40,
    "correct": 38,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.009153419733047485
  },
  "task1312_amazonreview_polarity_classification": {
    "count": 40,
    "correct": 40,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0003770291805267334
  },
  "task1313_amazonreview_polarity_classification": {
    "count": 40,
    "correct": 37,
    "accuracy": 0.925,
    "avg_semantic_distance": 0.011871118843555451
  },
  "task1333_check_validity_date_ddmmyyyy": {
    "count": 3,
    "correct": 2,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.05902034044265747
    },

"task1336_peixian_equity_evaluation_corpus_gender_classifier": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.000503474473953247
    },

"task1338_peixian_equity_evaluation_corpus_sentiment_classifier"
: {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0014616161584854125
    },
    "task133_winowhy_reason_plausibility_detection": {
      "count": 40,
      "correct": 21,
      "accuracy": 0.525,
      "avg_semantic_distance": 0.060255876183509825
    },
    "task1341_msr_text_classification": {
      "count": 3,
      "correct": 1,
      "accuracy": 0.3333333333333333,
      "avg_semantic_distance": 0.0777214765548706
    },
    "task1344_glue_entailment_classification": {
      "count": 40,
      "correct": 22,
      "accuracy": 0.55,
      "avg_semantic_distance": 0.07600716948509216
    },
    "task1346_glue_cola_grammatical_correctness_classification":
{
      "count": 20,
      "correct": 7,
      "accuracy": 0.35,
      "avg_semantic_distance": 0.08242697715759277
    },
    "task1347_glue_sts-b_similarity_classification": {
      "count": 22,
      "correct": 9,
      "accuracy": 0.4090909090909091,
      "avg_semantic_distance": 0.13795675201849503
    },
    "task1354_sent_comp_classification": {
      "count": 19,
      "correct": 16,
      "accuracy": 0.8421052631578947,
      "avg_semantic_distance": 0.06279630410043817
    },
    "task1361_movierationales_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.004514932632446289
    },
    "task1366_healthfact_classification": {
      "count": 1,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.1837763786315918
    },
    "task137_detoxifying-lms_classification_toxicity": {
      "count": 8,
      "correct": 4,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.03803486377000809
    },
"task1384_deal_or_no_dialog_classification": {
  "count": 40,
  "correct": 30,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.05581573247909546
},
"task1385_anli_r1_entailment": {
  "count": 19,
  "correct": 5,
  "accuracy": 0.2631578947368421,
  "avg_semantic_distance": 0.13274038779108147
},
"task1386_anli_r2_entailment": {
  "count": 19,
  "correct": 5,
  "accuracy": 0.2631578947368421,
  "avg_semantic_distance": 0.13943312042637876
},
"task1387_anli_r3_entailment": {
  "count": 22,
  "correct": 7,
  "accuracy": 0.3181818181818182,
  "avg_semantic_distance": 0.13437767733227124
},
"task1388_cb_entailment": {
  "count": 5,
  "correct": 4,
  "accuracy": 0.8,
  "avg_semantic_distance": 0.09968929290771485
},
"task138_detoxifying-lms_classification_fluency": {
  "count": 8,
  "correct": 4,
  "accuracy": 0.5,
  "avg_semantic_distance": 0.0374109148979187
},
"task1390_wscfixed_coreference": {
  "count": 12,
  "correct": 9,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.03845321635405222
},
"task1393_superglue_copa_text_completion": {
  "count": 9,
  "correct": 3,
  "accuracy": 0.3333333333333333,
  "avg_semantic_distance": 0.07303161091274685
},
"task139_detoxifying-lms_classification_topicality": {
  "count": 8,
  "correct": 3,
  "accuracy": 0.375,
  "avg_semantic_distance": 0.040319472551345825
},
"task1403_check_validity_date_mmddyyyy": {
  "count": 3,
  "correct": 3,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.013098716735839844
},
"task140_detoxifying-lms_classification_style": {
  "count": 8,
  "correct": 5,
  "accuracy": 0.625,
  "avg_semantic_distance": 0.03412114083766937
},
"task1418_bless_semantic_relation_classification": {
  "count": 20,
  "correct": 10,
  "accuracy": 0.5,
  "avg_semantic_distance": 0.14004340171813964
},
"task1429_evalution_semantic_relation_classification": {
  "count": 15,
  "correct": 6,
  "accuracy": 0.4,
  "avg_semantic_distance": 0.18393715620040893
},
"task1434_head_qa_classification": {
  "count": 38,
  "correct": 28,
  "accuracy": 0.7368421052631579,
  "avg_semantic_distance": 0.07099909217734086
},
"task145_afs_argument_similarity_death_penalty": {
  "count": 31,
  "correct": 23,
  "accuracy": 0.7419354838709677,
  "avg_semantic_distance": 0.04520598342341761
},
"task146_afs_argument_similarity_gun_control": {
  "count": 35,
  "correct": 29,
  "accuracy": 0.8285714285714286,
  "avg_semantic_distance": 0.03605343103408813
},
"task147_afs_argument_similarity_gay_marriage": {
  "count": 29,
  "correct": 24,
  "accuracy": 0.8275862068965517,
  "avg_semantic_distance": 0.03536061993960676
},
"task1488_sarcasmdetection_headline_classification": {
  "count": 10,
  "correct": 5,
  "accuracy": 0.5,
  "avg_semantic_distance": 0.06324649453163148
},
"task1489_sarcasmdetection_tweet_classification": {
  "count": 2,
  "correct": 2,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.014469772577285767
},
"task148_afs_argument_quality_gay_marriage": {
  "count": 21,
  "correct": 21,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.00854187636148362
},
"task1495_adverse_drug_event_classification": {
  "count": 21,
  "correct": 19,
  "accuracy": 0.9047619047619048,
  "avg_semantic_distance": 0.013434540657770066
},
"task149_afs_argument_quality_death_penalty": {
  "count": 21,
  "correct": 17,
  "accuracy": 0.8095238095238095,
  "avg_semantic_distance": 0.032450275761740546
},
"task1500_dstc3_classification": {
  "count": 9,
  "correct": 6,
  "accuracy": 0.6666666666666666,
  "avg_semantic_distance": 0.1316569646199544
},
"task1502_hatexplain_classification": {
  "count": 29,
  "correct": 16,
  "accuracy": 0.5517241379310345,
  "avg_semantic_distance": 0.10526388061457667
},
"task1503_hatexplain_classification": {
  "count": 17,
  "correct": 9,
  "accuracy": 0.5294117647058824,
  "avg_semantic_distance": 0.11031510900048648
},
"task1505_root09_semantic_relation_classification": {
  "count": 20,
  "correct": 14,
  "accuracy": 0.7,
  "avg_semantic_distance": 0.08451927900314331
},
"task150_afs_argument_quality_gun_control": {
  "count": 21,
  "correct": 21,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.0013886548223949614
},
"task1517_limit_classfication": {
  "count": 21,
  "correct": 15,
  "accuracy": 0.7142857142857143,
  "avg_semantic_distance": 0.05852693035489037
},
"task1529_scitail1.1_classification": {
  "count": 38,
  "correct": 28,
  "accuracy": 0.7368421052631579,
  "avg_semantic_distance": 0.08590490410202428
},
"task1531_daily_dialog_type_classification": {
  "count": 10,
  "correct": 6,
  "accuracy": 0.6,
  "avg_semantic_distance": 0.08704204559326172
},
  "task1533_daily_dialog_formal_classification": {
    "count": 40,
    "correct": 25,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.05962281227111817
  },
  "task1534_daily_dialog_question_classification": {
    "count": 40,
    "correct": 33,
    "accuracy": 0.825,
    "avg_semantic_distance": 0.04707983136177063
  },
  "task1541_agnews_classification": {
    "count": 40,
    "correct": 33,
    "accuracy": 0.825,
    "avg_semantic_distance": 0.05098218768835068
  },
  "task1548_wiqa_binary_classification": {
    "count": 11,
    "correct": 0,
    "accuracy": 0.0,
    "avg_semantic_distance": 0.0
  },
  "task1554_scitail_classification": {
    "count": 40,
    "correct": 30,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.08042175918817521
  },
  "task1559_blimp_binary_classification": {
    "count": 20,
    "correct": 8,
    "accuracy": 0.4,
    "avg_semantic_distance": 0.06634940505027771
  },
"task1560_blimp_binary_classification": {
  "count": 20,
  "correct": 14,
  "accuracy": 0.7,
  "avg_semantic_distance": 0.04108897149562836
},
"task1565_triviaqa_classification": {
  "count": 2,
  "correct": 0,
  "accuracy": 0.0,
  "avg_semantic_distance": 0.08445465564727783
},
"task1568_propara_classification": {
  "count": 4,
  "correct": 0,
  "accuracy": 0.0,
  "avg_semantic_distance": 0.0
},
"task156_codah_classification_adversarial": {
  "count": 37,
  "correct": 15,
  "accuracy": 0.40540540540540543,
  "avg_semantic_distance": 0.056293584205008844
},
"task1573_samsum_classification": {
  "count": 7,
  "correct": 3,
  "accuracy": 0.42857142857142855,
  "avg_semantic_distance": 0.05297814096723284
},
"task1583_bless_meronym_classification": {
  "count": 20,
  "correct": 16,
  "accuracy": 0.8,
  "avg_semantic_distance": 0.04222982823848724
},
"task1584_evalution_meronym_classification": {
  "count": 20,
  "correct": 14,
  "accuracy": 0.7,
  "avg_semantic_distance": 0.05955661535263061
},
"task1599_smcalflow_classification": {
  "count": 21,
  "correct": 20,
  "accuracy": 0.9523809523809523,
  "avg_semantic_distance": 0.008655573640550886
},
"task1604_ethos_text_classification": {
  "count": 19,
  "correct": 12,
  "accuracy": 0.631578947368421,
  "avg_semantic_distance": 0.05275435196725946
},
"task1605_ethos_text_classification": {
  "count": 6,
  "correct": 5,
  "accuracy": 0.8333333333333334,
  "avg_semantic_distance": 0.049790640672047935
},
"task1606_ethos_text_classification": {
  "count": 4,
  "correct": 2,
  "accuracy": 0.5,
  "avg_semantic_distance": 0.03759557008743286
},
"task1607_ethos_text_classification": {
  "count": 4,
  "correct": 3,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.05843760073184967
},
    "task1612_sick_label_classification": {
      "count": 21,
      "correct": 18,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.038660551820482524
    },
    "task1624_disfl_qa_question_yesno_classification": {
      "count": 19,
      "correct": 9,
      "accuracy": 0.47368421052631576,
      "avg_semantic_distance": 0.08224500480451082
    },

"task1640_aqa1.0_answerable_unanswerable_question_classification
": {
      "count": 21,
      "correct": 13,
      "accuracy": 0.6190476190476191,
      "avg_semantic_distance": 0.047402799129486084
    },

"task1645_medical_question_pair_dataset_text_classification": {
      "count": 40,
      "correct": 23,
      "accuracy": 0.575,
      "avg_semantic_distance": 0.05083464533090591
    },
    "task1661_super_glue_classification": {
      "count": 24,
      "correct": 14,
      "accuracy": 0.5833333333333334,
      "avg_semantic_distance": 0.08025785038868587
    },
    "task1705_ljspeech_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.009256541728973389
    },
    "task1706_ljspeech_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.010331302881240845
    },
    "task1712_poki_classification": {
      "count": 40,
      "correct": 32,
      "accuracy": 0.8,
      "avg_semantic_distance": 0.06582510471343994
    },
    "task1727_wiqa_what_is_the_effect": {
      "count": 25,
      "correct": 10,
      "accuracy": 0.4,
      "avg_semantic_distance": 0.09781059026718139
    },
    "task195_sentiment140_classification": {
      "count": 20,
      "correct": 17,
      "accuracy": 0.85,
      "avg_semantic_distance": 0.024762678146362304
    },
    "task196_sentiment140_answer_generation": {
      "count": 20,
      "correct": 17,
      "accuracy": 0.85,
      "avg_semantic_distance": 0.03559066355228424
    },
    "task211_logic2text_classification": {
      "count": 40,
      "correct": 36,
      "accuracy": 0.9,
      "avg_semantic_distance": 0.01430000513792038
    },
    "task212_logic2text_classification": {
      "count": 22,
      "correct": 22,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0027417242527008057
    },
    "task220_rocstories_title_classification": {
      "count": 29,
      "correct": 15,
      "accuracy": 0.5172413793103449,
      "avg_semantic_distance": 0.07088369131088257
    },
    "task226_english_language_answer_relevance_classification":
{
      "count": 6,
      "correct": 4,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.06840976079305013
    },
    "task227_clariq_classification": {
      "count": 20,
      "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.011829984188079835
    },
    "task232_iirc_link_number_classification": {
      "count": 20,
      "correct": 12,
      "accuracy": 0.6,
      "avg_semantic_distance": 0.06624206006526948
    },
    "task233_iirc_link_exists_classification": {
      "count": 20,
      "correct": 13,
      "accuracy": 0.65,
      "avg_semantic_distance": 0.06292637884616852
    },
    "task242_tweetqa_classification": {
      "count": 40,
      "correct": 38,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.014248554408550263
    },
    "task248_dream_classification": {
      "count": 12,
      "correct": 7,
      "accuracy": 0.5833333333333334,
      "avg_semantic_distance": 0.1288426419099172
    },
    "task274_overruling_legal_classification": {
      "count": 27,
      "correct": 25,
      "accuracy": 0.9259259259259259,
      "avg_semantic_distance": 0.01222471175370393
    },
    "task276_enhanced_wsc_classification": {
      "count": 26,
      "correct": 9,
      "accuracy": 0.34615384615384615,
      "avg_semantic_distance": 0.1615831943658682
    },
    "task279_stereoset_classification_stereotype": {
      "count": 20,
      "correct": 11,
      "accuracy": 0.55,
      "avg_semantic_distance": 0.08663694262504577
    },
    "task280_stereoset_classification_stereotype_type": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
    "avg_semantic_distance": 0.01659746113277617
  },
  "task284_imdb_classification": {
    "count": 22,
    "correct": 20,
    "accuracy": 0.9090909090909091,
    "avg_semantic_distance": 0.018154436891729183
  },
  "task285_imdb_answer_generation": {
    "count": 21,
    "correct": 18,
    "accuracy": 0.8571428571428571,
    "avg_semantic_distance": 0.037247697512308754
  },
  "task290_tellmewhy_question_answerability": {
    "count": 40,
    "correct": 23,
    "accuracy": 0.575,
    "avg_semantic_distance": 0.05248638540506363
  },
  "task296_storycloze_correct_end_classification": {
    "count": 22,
    "correct": 7,
    "accuracy": 0.3181818181818182,
    "avg_semantic_distance": 0.07397788492116061
  },
  "task297_storycloze_incorrect_end_classification": {
    "count": 22,
    "correct": 15,
    "accuracy": 0.6818181818181818,
    "avg_semantic_distance": 0.067415944554589
  },
  "task298_storycloze_correct_end_classification": {
    "count": 39,
    "correct": 19,
    "accuracy": 0.48717948717948717,
    "avg_semantic_distance": 0.07778637072978875
  },
  "task310_race_classification": {
    "count": 1,
    "correct": 0,
    "accuracy": 0.0,
    "avg_semantic_distance": 0.08375787734985352
  },
  "task316_crows-pairs_classification_stereotype": {
    "count": 20,
    "correct": 7,
    "accuracy": 0.35,
    "avg_semantic_distance": 0.05643531680107117
  },
  "task317_crows-pairs_classification_stereotype_type": {
    "count": 20,
    "correct": 14,
    "accuracy": 0.7,
    "avg_semantic_distance": 0.08008931875228882
  },
  "task318_stereoset_classification_gender": {
    "count": 15,
    "correct": 11,
    "accuracy": 0.7333333333333333,
    "avg_semantic_distance": 0.08141409158706665
  },
  "task319_stereoset_classification_profession": {
    "count": 21,
    "correct": 12,
    "accuracy": 0.5714285714285714,
    "avg_semantic_distance": 0.07296123107274373
  },
  "task320_stereoset_classification_race": {
    "count": 20,
    "correct": 16,
    "accuracy": 0.8,
      "avg_semantic_distance": 0.049490532279014586
    },
    "task321_stereoset_classification_religion": {
      "count": 4,
      "correct": 3,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.06870675086975098
    },
    "task322_jigsaw_classification_threat": {
      "count": 40,
      "correct": 39,
      "accuracy": 0.975,
      "avg_semantic_distance": 0.010635526478290558
    },
    "task323_jigsaw_classification_sexually_explicit": {
      "count": 40,
      "correct": 29,
      "accuracy": 0.725,
      "avg_semantic_distance": 0.032851383090019226
    },
    "task324_jigsaw_classification_disagree": {
      "count": 7,
      "correct": 6,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.05993264062064035
    },
    "task325_jigsaw_classification_identity_attack": {
      "count": 40,
      "correct": 33,
      "accuracy": 0.825,
      "avg_semantic_distance": 0.016598975658416747
    },
    "task326_jigsaw_classification_obscene": {
      "count": 40,
      "correct": 26,
      "accuracy": 0.65,
      "avg_semantic_distance": 0.05687294155359268
    },
    "task327_jigsaw_classification_toxic": {
      "count": 37,
      "correct": 34,
      "accuracy": 0.918918918918919,
      "avg_semantic_distance": 0.014244951106406547
    },
    "task328_jigsaw_classification_insult": {
      "count": 40,
      "correct": 34,
      "accuracy": 0.85,
      "avg_semantic_distance": 0.02431126981973648
    },
    "task329_gap_classification": {
      "count": 21,
      "correct": 6,
      "accuracy": 0.2857142857142857,
      "avg_semantic_distance": 0.115785056636447
    },
    "task333_hateeval_classification_hate_en": {
      "count": 33,
      "correct": 27,
      "accuracy": 0.8181818181818182,
      "avg_semantic_distance": 0.03343028010744037
    },
    "task335_hateeval_classification_aggresive_en": {
      "count": 28,
      "correct": 11,
      "accuracy": 0.39285714285714285,
      "avg_semantic_distance": 0.06807514812265124
    },
    "task337_hateeval_classification_individual_en": {
      "count": 26,
      "correct": 23,
      "accuracy": 0.8846153846153846,
        "avg_semantic_distance": 0.024406362038392287
      },
      "task340_winomt_classification_gender_pro": {
        "count": 20,
        "correct": 20,
        "accuracy": 1.0,
        "avg_semantic_distance": 0.0028349846601486207
      },
      "task341_winomt_classification_gender_anti": {
        "count": 20,
        "correct": 20,
        "accuracy": 1.0,
        "avg_semantic_distance": 0.0021029263734817505
      },
      "task346_hybridqa_classification": {
        "count": 21,
        "correct": 6,
        "accuracy": 0.2857142857142857,
        "avg_semantic_distance": 0.056822512831006734
      },

"task349_squad2.0_answerable_unanswerable_question_classificatio
n": {
        "count": 21,
        "correct": 14,
        "accuracy": 0.6666666666666666,
        "avg_semantic_distance": 0.04619585900079636
      },
      "task350_winomt_classification_gender_identifiability_pro":
{
        "count": 20,
        "correct": 9,
        "accuracy": 0.45,
        "avg_semantic_distance": 0.05230997204780578
      },
      "task351_winomt_classification_gender_identifiability_anti":
{
      "count": 20,
      "correct": 12,
      "accuracy": 0.6,
      "avg_semantic_distance": 0.04925930798053742
    },
    "task353_casino_classification_negotiation_elicit_pref": {
      "count": 14,
      "correct": 10,
      "accuracy": 0.7142857142857143,
      "avg_semantic_distance": 0.05587678721972874
    },
    "task354_casino_classification_negotiation_no_need": {
      "count": 7,
      "correct": 5,
      "accuracy": 0.7142857142857143,
      "avg_semantic_distance": 0.06045009408678327
    },
    "task355_casino_classification_negotiation_other_need": {
      "count": 16,
      "correct": 11,
      "accuracy": 0.6875,
      "avg_semantic_distance": 0.0575757659971714
    },
    "task356_casino_classification_negotiation_self_need": {
      "count": 30,
      "correct": 21,
      "accuracy": 0.7,
      "avg_semantic_distance": 0.06025944550832112
    },
    "task357_casino_classification_negotiation_small_talk": {
      "count": 36,
      "correct": 28,
      "accuracy": 0.7777777777777778,
      "avg_semantic_distance": 0.051419887277815074
    },
    "task358_casino_classification_negotiation_uv_part": {
      "count": 4,
      "correct": 3,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.07354827225208282
    },
    "task359_casino_classification_negotiation_vouch_fair": {
      "count": 17,
      "correct": 9,
      "accuracy": 0.5294117647058824,
      "avg_semantic_distance": 0.06705757449654971
    },
    "task362_spolin_yesand_prompt_response_sub_classification":
{
      "count": 40,
      "correct": 13,
      "accuracy": 0.325,
      "avg_semantic_distance": 0.037712198495864865
    },
    "task363_sst2_polarity_classification": {
      "count": 21,
      "correct": 18,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.05039086795988537
    },
    "task364_regard_social_impact_classification": {
      "count": 5,
      "correct": 3,
      "accuracy": 0.6,
      "avg_semantic_distance": 0.12232996225357055
    },
    "task375_classify_type_of_sentence_in_debate": {
      "count": 8,
      "correct": 6,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.10511460900306702
    },
    "task379_agnews_topic_classification": {
      "count": 40,
      "correct": 35,
      "accuracy": 0.875,
      "avg_semantic_distance": 0.032955878973007204
    },
    "task383_matres_classification": {
      "count": 32,
      "correct": 27,
      "accuracy": 0.84375,
      "avg_semantic_distance": 0.03687131777405739
    },
    "task384_socialiqa_question_classification": {
      "count": 21,
      "correct": 16,
      "accuracy": 0.7619047619047619,
      "avg_semantic_distance": 0.02713345345996675
    },
    "task386_semeval_2018_task3_irony_detection": {
      "count": 21,
      "correct": 17,
      "accuracy": 0.8095238095238095,
      "avg_semantic_distance": 0.06156925644193377
    },
    "task387_semeval_2018_task3_irony_classification": {
      "count": 21,
      "correct": 13,
      "accuracy": 0.6190476190476191,
      "avg_semantic_distance": 0.11973357768285842
    },
    "task397_semeval_2018_task1_tweet_anger_detection": {
      "count": 20,
      "correct": 17,
      "accuracy": 0.85,
      "avg_semantic_distance": 0.0334204375743866
    },
    "task398_semeval_2018_task1_tweet_joy_detection": {
```

```
    "count": 20,
    "correct": 15,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.029283204674720766
  },
  "task399_semeval_2018_task1_tweet_sadness_detection": {
    "count": 20,
    "correct": 15,
    "accuracy": 0.75,
    "avg_semantic_distance": 0.0407205194234848
  },
  "task456_matres_intention_classification": {
    "count": 36,
    "correct": 32,
    "accuracy": 0.8888888888888888,
    "avg_semantic_distance": 0.02245947884188758
  },
  "task457_matres_conditional_classification": {
    "count": 12,
    "correct": 12,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0038242091735204062
  },
  "task458_matres_negation_classification": {
    "count": 14,
    "correct": 11,
    "accuracy": 0.7857142857142857,
    "avg_semantic_distance": 0.048503675631114414
  },
  "task459_matres_static_classification": {
    "count": 2,
    "correct": 2,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.002063363790512085
  },
  "task462_qasper_classification": {
    "count": 24,
    "correct": 12,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.13438450545072556
  },
  "task472_haspart_classification": {
    "count": 20,
    "correct": 13,
    "accuracy": 0.65,
    "avg_semantic_distance": 0.06449390649795532
  },
  "task475_yelp_polarity_classification": {
    "count": 40,
    "correct": 38,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.009599690139293671
  },
  "task476_cls_english_books_classification": {
    "count": 25,
    "correct": 25,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.005852947235107422
  },
  "task477_cls_english_dvd_classification": {
    "count": 26,
    "correct": 25,
    "accuracy": 0.9615384615384616,
    "avg_semantic_distance": 0.013045347653902493
  },
  "task478_cls_english_music_classification": {
    "count": 28,
    "correct": 26,
    "accuracy": 0.9285714285714286,
    "avg_semantic_distance": 0.021561665194375173
  },
  "task493_review_polarity_classification": {
    "count": 40,
    "correct": 39,
    "accuracy": 0.975,
    "avg_semantic_distance": 0.002506695687708435
  },
  "task494_review_polarity_answer_generation": {
    "count": 40,
    "correct": 37,
    "accuracy": 0.925,
    "avg_semantic_distance": 0.014130321145057679
  },
  "task495_semeval_headline_classification": {
    "count": 20,
    "correct": 10,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.06439176797866822
  },
  "task512_twitter_emotion_classification": {
    "count": 21,
    "correct": 14,
    "accuracy": 0.6666666666666666,
    "avg_semantic_distance": 0.06751241570427305
  },
  "task514_argument_consequence_classification": {
    "count": 3,
    "correct": 2,
    "accuracy": 0.6666666666666666,
    "avg_semantic_distance": 0.05147528648376465
  },
  "task517_emo_classify_emotion_of_dialogue": {
    "count": 21,
    "correct": 18,
    "accuracy": 0.8571428571428571,
    "avg_semantic_distance": 0.0455610610189892
  },
  "task518_emo_different_dialogue_emotions": {
      "count": 26,
      "correct": 15,
      "accuracy": 0.5769230769230769,
      "avg_semantic_distance": 0.06808157150561993
    },
    "task521_trivia_question_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.017913809844425747
    },
    "task564_discofuse_classification": {
      "count": 19,
      "correct": 5,
      "accuracy": 0.2631578947368421,
      "avg_semantic_distance": 0.12476744463569239
    },
    "task566_circa_classification": {
      "count": 20,
      "correct": 15,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.04753294289112091
    },
    "task573_air_dialogue_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.018888748827434722
    },
    "task575_air_dialogue_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.005693188309669495
    },
    "task577_curiosity_dialogs_classification": {
      "count": 23,
      "correct": 23,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.0032043742096942406
    },
    "task579_socialiqa_classification": {
      "count": 40,
      "correct": 23,
      "accuracy": 0.575,
      "avg_semantic_distance": 0.07603603452444077
    },
    "task585_preposition_classification": {
      "count": 18,
      "correct": 4,
      "accuracy": 0.2222222222222222,
      "avg_semantic_distance": 0.14275875025325352
    },
    "task586_amazonfood_polarity_classification": {
      "count": 40,
      "correct": 37,
      "accuracy": 0.925,
      "avg_semantic_distance": 0.014340519905090332
    },
    "task587_amazonfood_polarity_correction_classification": {
      "count": 40,
      "correct": 35,
      "accuracy": 0.875,
      "avg_semantic_distance": 0.013649210333824158
    },
    "task588_amazonfood_rating_classification": {
      "count": 40,
      "correct": 28,
      "accuracy": 0.7,
      "avg_semantic_distance": 0.08290962874889374
    },
    "task590_amazonfood_summary_correction_classification": {
      "count": 40,
      "correct": 17,
      "accuracy": 0.425,
      "avg_semantic_distance": 0.05877882242202759
    },
    "task607_sbic_intentional_offense_binary_classification": {
      "count": 21,
      "correct": 13,
      "accuracy": 0.6190476190476191,
      "avg_semantic_distance": 0.062255115736098515
    },
    "task608_sbic_sexual_offense_binary_classification": {
      "count": 21,
      "correct": 18,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.03208814632324945
    },
    "task609_sbic_potentially_offense_binary_classification": {
      "count": 21,
      "correct": 16,
      "accuracy": 0.7619047619047619,
      "avg_semantic_distance": 0.044448290552411764
    },
    "task616_cola_classification": {
      "count": 20,
      "correct": 12,
      "accuracy": 0.6,
      "avg_semantic_distance": 0.06761164963245392
    },
    "task623_ohsumed_yes_no_answer_generation": {
      "count": 3,
      "correct": 3,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.033493220806121826
    },
    "task625_xlwic_true_or_false_answer_generation": {
```

```
  "count": 2,
  "correct": 2,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.05840137600898743
},
"task638_multi_woz_classification": {
  "count": 21,
  "correct": 13,
  "accuracy": 0.6190476190476191,
  "avg_semantic_distance": 0.06350085848853701
},
"task673_google_wellformed_query_classification": {
  "count": 20,
  "correct": 15,
  "accuracy": 0.75,
  "avg_semantic_distance": 0.03876414895057678
},
"task679_hope_edi_english_text_classification": {
  "count": 26,
  "correct": 20,
  "accuracy": 0.7692307692307693,
  "avg_semantic_distance": 0.025666608260228083
},
"task681_hope_edi_malayalam_text_classification": {
  "count": 22,
  "correct": 18,
  "accuracy": 0.8181818181818182,
  "avg_semantic_distance": 0.03936533765359358
},
"task682_online_privacy_policy_text_classification": {
  "count": 13,
  "correct": 13,
  "accuracy": 1.0,
  "avg_semantic_distance": 0.023420980343451865
},
"task685_mmmlu_answer_generation_clinical_knowledge": {
    "count": 3,
    "correct": 0,
    "accuracy": 0.0,
    "avg_semantic_distance": 0.12107515335083008
  },
  "task698_mmmlu_answer_generation_global_facts": {
    "count": 2,
    "correct": 1,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.10824224352836609
  },
  "task721_mmmlu_answer_generation_medical_genetics": {
    "count": 2,
    "correct": 1,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.09285563230514526
  },
  "task738_perspectrum_classification": {
    "count": 21,
    "correct": 14,
    "accuracy": 0.6666666666666666,
    "avg_semantic_distance": 0.0931341704868135
  },
  "task746_yelp_restaurant_review_classification": {
    "count": 16,
    "correct": 16,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0020403414964675903
  },
  "task761_app_review_classification": {
    "count": 6,
    "correct": 6,
    "accuracy": 1.0,
    "avg_semantic_distance": 0.0006819566090901693
  },
  "task766_craigslist_bargains_classification": {
      "count": 3,
      "correct": 1,
      "accuracy": 0.3333333333333333,
      "avg_semantic_distance": 0.12009219328562419
    },
    "task767_craigslist_bargains_classification": {
      "count": 19,
      "correct": 18,
      "accuracy": 0.9473684210526315,
      "avg_semantic_distance": 0.02408228736174734
    },
    "task819_pec_sentiment_classification": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.003094404935836792
    },
    "task823_peixian-rtgender_sentiment_analysis": {
      "count": 40,
      "correct": 20,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.09653442651033402
    },
    "task827_copa_commonsense_reasoning": {
      "count": 19,
      "correct": 5,
      "accuracy": 0.2631578947368421,
      "avg_semantic_distance": 0.10084295586535805
    },
    "task828_copa_commonsense_cause_effect": {
      "count": 19,
      "correct": 10,
      "accuracy": 0.5263157894736842,
      "avg_semantic_distance": 0.07701548777128521
    },
    "task833_poem_sentiment_classification": {
      "count": 5,
      "correct": 4,
      "accuracy": 0.8,
      "avg_semantic_distance": 0.015591752529144288
    },
    "task834_mathdataset_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.002137395739555359
    },
    "task843_financial_phrasebank_classification": {
      "count": 21,
      "correct": 19,
      "accuracy": 0.9047619047619048,
      "avg_semantic_distance": 0.01857284704844157
    },
    "task844_financial_phrasebank_classification": {
      "count": 24,
      "correct": 18,
      "accuracy": 0.75,
      "avg_semantic_distance": 0.03269439438978831
    },
    "task846_pubmedqa_classification": {
      "count": 35,
      "correct": 30,
      "accuracy": 0.8571428571428571,
      "avg_semantic_distance": 0.032298251560756136
    },
    "task848_pubmedqa_classification": {
      "count": 16,
      "correct": 10,
      "accuracy": 0.625,
      "avg_semantic_distance": 0.07852806895971298
    },
    "task854_hippocorpus_classification": {
      "count": 1,
      "correct": 1,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.13010764122009277
    },
    "task855_conv_ai_2_classification": {
      "count": 2,
      "correct": 1,
      "accuracy": 0.5,
      "avg_semantic_distance": 0.0911327600479126
    },
    "task856_conv_ai_2_classification": {
      "count": 3,
      "correct": 2,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.06695004304250081
    },
    "task875_emotion_classification": {
      "count": 21,
      "correct": 14,
      "accuracy": 0.6666666666666666,
      "avg_semantic_distance": 0.08838629722595215
    },
    "task879_schema_guided_dstc8_classification": {
      "count": 20,
      "correct": 20,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.00045125484466552735
    },
    "task880_schema_guided_dstc8_classification": {
      "count": 20,
      "correct": 19,
      "accuracy": 0.95,
      "avg_semantic_distance": 0.032116946578025815
    },
    "task888_reviews_classification": {
      "count": 10,
      "correct": 9,
      "accuracy": 0.9,
      "avg_semantic_distance": 0.017439407110214234
    },
    "task890_gcwd_classification": {
      "count": 4,
      "correct": 1,
      "accuracy": 0.25,
      "avg_semantic_distance": 0.1856546252965927
    },
    "task902_deceptive_opinion_spam_classification": {
      "count": 21,
      "correct": 20,
      "accuracy": 0.9523809523809523,
      "avg_semantic_distance": 0.009843462989443824
    },
    "task903_deceptive_opinion_spam_classification": {
      "count": 21,
      "correct": 13,
      "accuracy": 0.6190476190476191,
      "avg_semantic_distance": 0.03511201767694382
    },
    "task907_dialogre_identify_relationships": {
      "count": 2,
      "correct": 0,
      "accuracy": 0.0,
      "avg_semantic_distance": 0.11374858021736145
    },
    "task908_dialogre_identify_familial_relationships": {
      "count": 2,
      "correct": 2,
      "accuracy": 1.0,
      "avg_semantic_distance": 0.038915663957595825
    },
    "task921_code_x_glue_information_retreival": {
```

```
    "count": 8,
    "correct": 5,
    "accuracy": 0.625,
    "avg_semantic_distance": 0.12098922580480576
  },
  "task923_event2mind_classifier": {
    "count": 20,
    "correct": 11,
    "accuracy": 0.55,
    "avg_semantic_distance": 0.09871419072151184
  },
  "task925_coached_conv_pref_classifier": {
    "count": 10,
    "correct": 7,
    "accuracy": 0.7,
    "avg_semantic_distance": 0.06612423658370972
  },
  "task929_products_reviews_classification": {
    "count": 40,
    "correct": 38,
    "accuracy": 0.95,
    "avg_semantic_distance": 0.011376681923866271
  },
  "task935_defeasible_nli_atomic_classification": {
    "count": 20,
    "correct": 12,
    "accuracy": 0.6,
    "avg_semantic_distance": 0.03793564140796661
  },
  "task937_defeasible_nli_social_classification": {
    "count": 20,
    "correct": 10,
    "accuracy": 0.5,
    "avg_semantic_distance": 0.041941669583320615
  },
  "task966_ruletaker_fact_checking_based_on_given_context": {
      "count": 5,
      "correct": 2,
      "accuracy": 0.4,
      "avg_semantic_distance": 0.06376264095306397
    }
  },
  "note": "Detailed results saved in validation_results/
directory (260 task files)"
}
```