future internet

MDPI

Article

Testing Stimulus Equivalence in Transformer-Based Agents

Alexis Carrillo and Moisés Betancort *

Departamento de Psicología Clínica, Psicobiología y Metodología, Campus de Guajara, Universidad de La Laguna, Apartado 456, 38200 San Cristóbal de La Laguna, Spain
* Correspondence: moibemo@ull.edu.es

Abstract: This study investigates the ability of transformer-based models (TBMs) to form stimulus equivalence (SE) classes. We employ BERT and GPT as TBM agents in SE tasks, evaluating their performance across training structures (linear series, one-to-many and many-to-one) and relation types (select–reject, select-only). Our findings demonstrate that both models performed above mastery criterion in the baseline phase across all simulations (n = 12). However, they exhibit limited success in reflexivity, transitivity, and symmetry tests. Notably, both models achieved success only in the linear series structure with select–reject relations, failing in one-to-many and many-to-one structures, and all select-only conditions. These results suggest that TBM may be forming decision rules based on learned discriminations and reject relations, rather than responding according to equivalence class formation. The absence of reject relations appears to influence their responses and the occurrence of hallucinations. This research highlights the potential of SE simulations for: (a) comparative analysis of learning mechanisms, (b) explainability techniques for TBM decision-making, and (c) TBM benchmarking independent of pre-training or fine-tuning. Future investigations can explore upscaling simulations and utilize SE tasks within a reinforcement learning framework.

Keywords: stimulus equivalence; transformers; BERT; GPT; matching to sample; reject control; training structures

Citation: Carrillo, A.; Betancort, M. Testing Stimulus Equivalence in Transformer-Based Agents. *Future Internet* **2024**, *16*, 289. https://doi.org/10.3390/fi16080289

Academic Editor: Paolo Bellavista

Received: 29 June 2024
Revised: 2 August 2024
Accepted: 6 August 2024
Published: 9 August 2024

1. Introduction

Stimulus equivalence (SE) is a behavioral phenomenon characterized by the emergence of novel stimulus control without explicit training [1–3]. This ability is fundamental to human language and cognition [4]. Understanding the mechanisms underlying this process is crucial for developing effective interventions in areas such as education, therapy, and artificial intelligence. While extensively studied in humans and animals [5], the computational modelling of stimulus equivalence [6], particularly using complex architectures like transformers, remains relatively under-explored.

1.1. Stimulus Equivalence

SE is defined as responding in accordance with the features of reflexivity, symmetry, and transitivity [1]. Stimulus classes can be formed from any unrelated stimulus, arbitrarily assigned, and functionally related regardless of its physical properties [7]. If a stimulus controls one member of the class, it affects all members of an equivalent stimulus class [2]. In SE, functional properties can be transferred to other stimuli without explicit training among members of an equivalence class. A stimulus trained as a conditional stimulus can form new analytical units that have not been previously trained, in which it can either control other three-term units or in which its role changes to a discriminative stimulus under the control of other stimuli [1]. SE provides a framework in the experimental analysis of behaviour to study language, symbolic behaviour, and cognition.

The most common framework for training and evaluating SE is arbitrary matching-to-sample (MTS). The goal of this procedure is training conditional discriminations presenting a sample stimulus and two or more comparison stimuli. After training, individuals are

assessed for responding in accordance with SE by conducting reflexivity, symmetry, and transitivity tests [1,2,5,7,8].

An important line of research on SE formation is the conditions that could influence the emergence of equivalence classes [7,8]. One methodological aspect of SE research is the training structure (TS). TS determines which pairs of stimuli are used for baseline training and which pairs are used to test the properties of reflexivity, symmetry, and transitivity. Three basic patterns of TS are typically used: linear series (LS), many-to-one (MTO), and one-to-many (OTM) [5,8,9]. Figure 1 shows a graphical representation of train structures in a four-members (A, B, C, and D) equivalence class with their respective pairs of reflexivity, symmetry, and transitivity. In LS, stimuli are presented sequentially, forming a chain of conditional discriminations. The stimuli are arranged in a specific order, such as A-B-C-D, establishing a clear and sequential relations between each stimulus and its adjacent counterparts. In OTM, a single sample stimulus is associated with multiple comparison stimuli with the pairs A-B, A-C, and A-D. In MTO, multiple sample stimuli are associated with a single comparison stimulus, B-A, C-A, D-A, wherein stimuli B, C, and D serve as samples and are related to a common stimulus A.

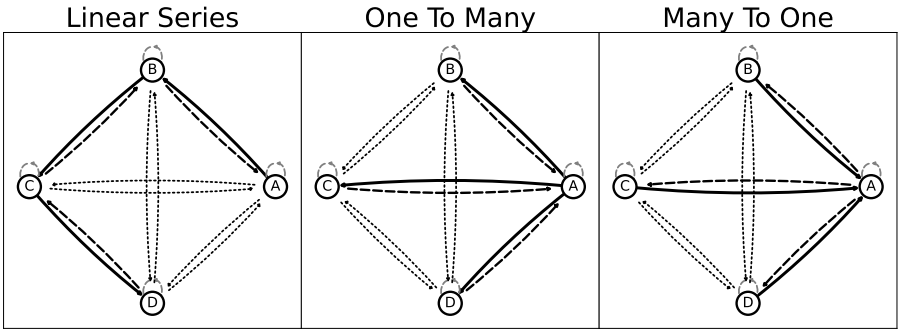

Figure 1. Training structures for members A, B, C, and D of an equivalence class. Baseline relations are shown in black solid arrows. Emergent relations for testing are reflexivity in dashed grey arrows; symmetry in black dashed arrows; and transitivity in black dotted arrows.

Another methodological consideration is the type of relation that is formed in MTS. In a typical MTS experiment, there is one sample stimulus and three comparison stimuli. During a baseline training trial, one of the comparison stimuli is reinforced, and the other are either punished or non-contingent. Hence, two types of relations between stimuli are formed: select type relations and reject type relations. Select type relations are established when the organism consistently responds to the correct comparison stimulus when presented with the sample stimulus, indicating which comparison stimuli are appropriate matches for the sample stimulus. Reject type relations are established when the organism consistently avoids choosing incorrect comparison stimuli when presented with the sample stimulus, indicating which comparison stimuli are not suitable matches. There is no clear evidence regarding whether SE requires select-only or also requires both select and reject relations. Carrigan and Sidman [10] suggest that select-only relations are essential, though different experiments yield varied results [11–14].

1.2. Related Work

SE experiments, often conducted on humans, can also be simulated using computational models [6]. Studies have been conducted on SE employing connectionist agents. These agents are variations of feed-forward artificial neural networks (ANNs) trained with different learning algorithms and applied to various SE tasks. In RELNET, ref. [15] a single hidden layer feed-forward network was used for MTS tasks, simulating derived relations under contextual control. A flaw existed in its sample-marking duplicator, which provided identical activation patterns for different trials. Tovar and Torres-Chávez [16] also implemented a three-layer feed-forward network with back-propagation learning. It focused on compound stimuli procedures with YES/NO responses, avoiding the need

for sample marking. Vernucio and Debert [17] presented a modified version of the Tovar and Torres-Chávez network. It used go/no-go responses during compound stimuli procedures with a single output unit for "go" responses. Emergent Virtual Analytics (EVA) [18] is a computational resource for simulating SE. It builds upon previous models and allows for an exploration of both theoretical and applied aspects of SE. Its strengths include the ability to accommodate various training protocols and offer deeper network architectures (four layers) to simulate more complex human behaviour like contextual control. Carrillo and Betancort [19] investigated the impact of TS on stimulus class formation, highlighted a critical limitation in previous computational models of stimulus equivalence, including RELNET, Tovar, and Ninnes' EVA, related to stimulus encoding and evaluation, and implemented multiple TSs instead of a single one and utilised four ANN architectures with varying complexity. Additionally, the authors proposed a new input encoding scheme for a more comprehensive evaluation of emergent relations. While full SE was not achieved, emergent properties of reflexivity and transitivity were observed in specific TS and ANN configurations.

Equivalence projective simulation (EPS) [20] and enhanced equivalence projective simulation (E-EPS) [21] models developed by Mofrad and colleagues employ a reinforcement learning framework with an episodic memory network to capture the dynamics of stimulus equivalence. While EPS relies on a training phase to shape the memory network and a subsequent test phase for evaluating derived relations, E-EPS incorporates a more dynamic approach by allowing the memory network to evolve during testing. Both models have demonstrated success in simulating key aspects of stimulus equivalence, including the formation of equivalence classes and the impact of different training structures.

While feed-forward ANNs have been used for simulating SE, their performance shows limitations in generalizing complex relationships within stimuli. This lead to a search for other alternatives for the agents. Feed-forward ANNs are part of a broad family of algorithms known as deep learning (DL). Variations in the number of layers, their interconnections, and the mathematical functions used, describe the architecture of DL models. Common DL architectures include deep neural networks, recurrent neural networks, convolutional neural networks, and transformers [22–24]. This research uses transformer-based models (TBMs) as agents.

1.3. Transformer-Based Models

Transformers [25] are neural networks that rely on self-attention mechanisms. This allows them to efficiently process both sequential and non-sequential data in parallel, focus on specific parts of the input, and learn relations between different parts, making it suited for tasks requiring textual understanding, such as machine translation and question answering. The attention mechanism allows them to identify relevant parts of the input simultaneously, regardless of their position. This makes them well-suited for tasks involving complex relationships within data, potentially leading to a more nuanced understanding of SE compared to feed-forward models.

Bidirectional Encoder Representations from Transformers (BERT) was designed for natural language understanding tasks [26]. It uses a training method called masked language model objective, which allows it to capture bidirectional context, making it suitable for understanding relations between words in both directions. BERT excels at tasks requiring deep contextual understanding, such as question answering, sentiment analysis, and named entity recognition. It is typically fine-tuned for specific tasks after pre-training on a large corpus.

Generative Pretrained Transformer (GPT) is designed for both language generation and understanding [27,28]. The GPT series uses an autoregressive training approach wherein the model predicts the next word in a sequence based on past context. GPT processes context unidirectionally, generating text based on preceding context. This makes it well-suited for text generation tasks like chatbots, content creation, and story generation. Additionally, GPT can perform few-shot learning, adapting to new tasks with minimal fine-tuning based on a few examples. BERT's bidirectional understanding may be advantageous in equivalence scenarios requiring nuanced contextual relationships. GPT,

with its generative capabilities, may excel in tasks requiring creative and contextually appropriate responses.

Although research has explored SE in fully connected feed-forward networks, to the best of our knowledge, we did not find studies that have tested SE on other deep learning architectures, specifically TBM. Understanding how humans learn and utilize symbols, as in language, is important for the field of natural language processing. If transformers can effectively model SE, it could provide insights into core language functions. This connection is relevant because transformers are the architecture for foundational language models (FLMs).

FLMs, also known as large language models, are trained on massive amounts of text data, enabling them to perform feats like generating human-quality text, translating languages, and writing various creative content. The extent of their abilities remains an open question. Proponents point to observations suggesting potential for advanced reasoning capabilities in FLMs [29,30] as Chain-of-Thought (CoT) prompting [31,32]. CoT prompting is a technique that breaks down reasoning into smaller steps, guiding FLMs through the problem-solving process. However, critics raise concerns about the true nature of these capabilities. Despite their success, there is ongoing debate about how well FLMs truly understand language [33–35] and limitations as the reversal curse [36], and hallucinations [37]. The reversal curse highlights limitations in current unidirectional models like GPT. While these models excel at learning forward relationships (A is B, B is C, therefore A is C), they struggle with tasks requiring reversing the reasoning process (B is A, C is B, or C is A). FLMs can generate outputs that are factually incorrect, nonsensical, or irrelevant to the prompt. These hallucinations can be misleading and require careful evaluation to avoid misinterpreting FLM capabilities. We refer to the models employed in our simulations as TBMs throughout this paper to differentiate them from larger, pre-trained FLMs.

Computational modeling offers a promising avenue for exploring the mechanisms underlying stimulus equivalence. This study investigates the capacity of TBMs to form equivalence classes. By comparing the performance of BERT and GPT models across different training conditions, we aim to shed light on how these architectures process and represent symbolic information. Specifically, we seek to: (1) Determine whether TBMs can form equivalence classes comparable to those observed in humans and animals. (2) Explore how different training structures and relation types influence TBM performance. By addressing these research questions, we aim to advance our understanding of how TBMs can process and represent symbolic information, contributing to the development of more sophisticated language models. SE is a tool for investigating the fundamental processes underlying the ability to relate and generalize across symbols. If TBMs demonstrate SE, this may suggest that they possess a more fundamental understanding of language beyond statistical patterns [33]. This could lead to the development of FLMs that not only process language but also truly grasp its underlying meaning and structure, opening doors for more advanced human–computer communication and language-based applications.

2. Materials and Methods

A total of 12 simulations were designed according to the combination of two transformer architectures, three train structures, and two relation type conditions.

2.1. Computational Agents Architecture

Computational agents have been built based on Andrej Karpathy's NanoGPT code, available on GitHub [38]. It is a minimal code implementation for a language model such as GPT, which includes a decoder block with a masked self-attention head. A modified version of NanoGPT was utilized for a second agent based on BERT [26], which consists of an encoder block with an unmasked self-attention head. Consequently, GPT can only search for information from tokens preceding the current token being processed, whereas in BERT, the complete sequence is accessible due to the absence of a filter or blocking of information with a mask. The architecture of the agents is presented in Figure 2, with

positional encoding token embedding combined, serving as input for the block. Apart from this divergence, these agents employ the same architecture.

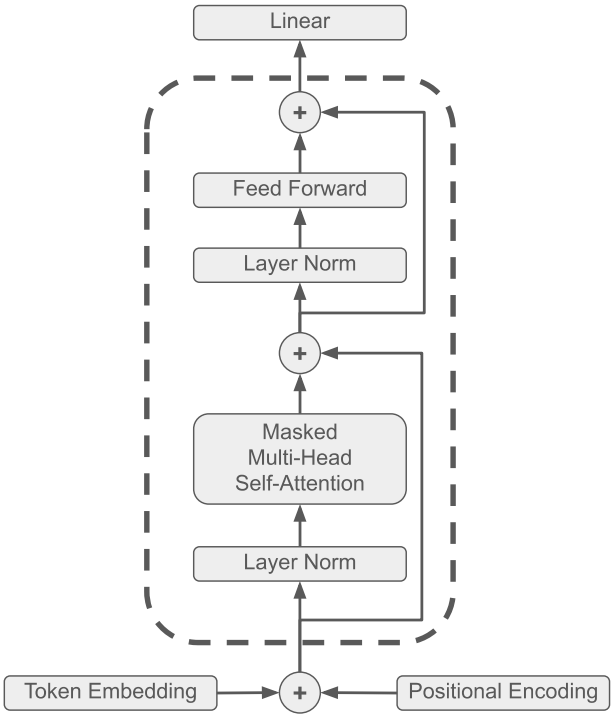

Figure 2. NanoGPT transformer architecture, as described by Karpathy.

The rationale behind the decision to use this code, and only these structures on an untrained model tested directly on trials, is to control unwanted effects of prior knowledge. We employed a minimal code implementation of BERT and GPT (10 million parameters) without pre-training to isolate the effects of SE learning on model performance. This approach directly evaluates the models' ability to learn relationships from the experimental trials. Pre-trained models introduce confounding variables as their weights reflect prior exposure to unknown information sets. Here, both models received identical training data, ensuring results solely depend on the SE task and not pre-existing knowledge embedded within pre-trained weights. This methodology mitigates potential biases arising from uncontrolled pre-training, allowing for a focused investigation of SE learning capabilities in TBMs.

To isolate the inherent capabilities of TBMs in acquiring stimulus equivalence, we intentionally excluded fine-tuning. This approach allowed us to observe how models develop internal representations without the potential influence of biases introduced through exposure to pre-trained data. Fine-tuning could lead to high performance based solely on memorized patterns or specific relations learned during this process, rather than a genuine understanding of equivalence class formation. By omitting fine-tuning, we aimed to uncover the core mechanisms underlying stimulus equivalence in these models.

2.2. Experimental Dataset Creation

Training baseline pairs followed the procedure of simultaneous MTS with one sample and three comparisons per trial. Baseline relations were designed to train four classes (C = 1, 2, 3, 4) with seven members (M = A, B, C, D, E, F, G) for a total of (M × C) 28 class member stimuli (A1, B1, C1, D1, A2, . . .G4). Trials creation is based on Carrillo and Betancort [19] steps, from the combination of the 28 stimuli in four positions (28^4 = 614,656), wherein are selected those trials in which no comparison is repeated and only one comparison is a valid pair with the sample, across baseline (n = 30,240), reflexibility (n = 35,280), symmetry (n = 30,240), and transitivity (n = 151,200) pairs, for a total of 246,960 trials. Six sets of trials were created from three train structures (LS, OTM, and MTO) and two relation

type conditions (select–reject, select-only). Figure 3 shows the train structures for the simulations. Response options are labeled with the letter O, an underscore character, and numbers one to three (O_1, O_2, O_3,) and represents the position of the comparison stimulus as if a human had to press one of three keys as response. A more detailed description of the train structures is presented in Appendix A.

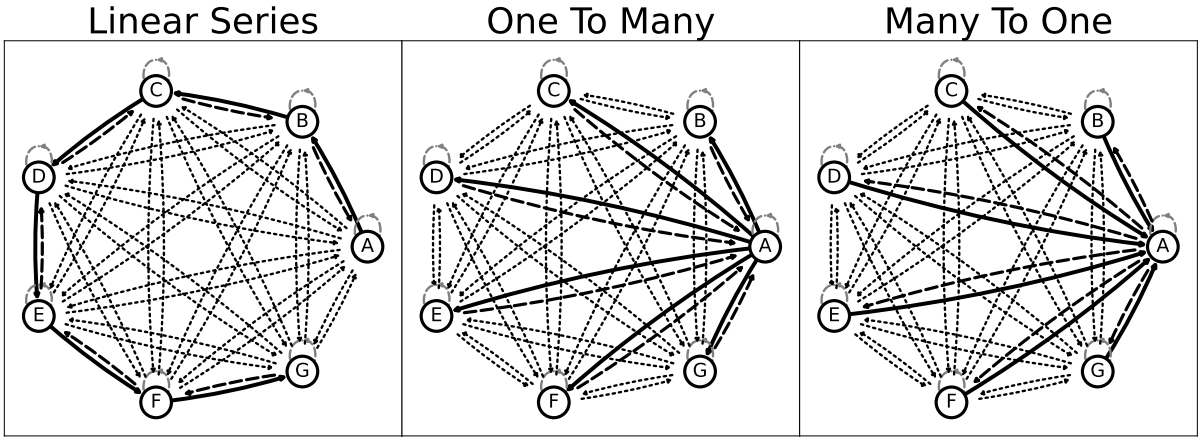

Figure 3. Experimental pairs according to train structures condition. Baseline relations are shown in black solid arrows. Emergent relations for testing are reflexivity in dashed grey arrows; symmetry in black dashed arrows; and transitivity in black dotted arrows.

The select–reject condition corresponds to the standard use of other class members as incorrect comparisons, with the subsequent simultaneous train of select–reject relations. A modification of the altered matching to sample procedure [14] is employed to exclusively train select relations during the train phase. For the select-only relation type condition, an additional set of $(M(C-1))$ 21 dummy stimuli is added, labeled with the letter "Z", an underscore character, and numbers from 11 to 31 (Z_11, Z_12, Z_13...). Dummy stimuli are used as incorrect comparisons on the baseline trials, in replacement of the members of other classes and diverge the establishment of reject relations from other class members stimuli, making it possible to test SE solely on select relations. Evaluation of reflexivity, symmetry, and transitivity uses members of other classes as comparisons, making it possible to compare the responses with the same trials in both relation type conditions. Select-only trials diverge from the altered matching to sample procedure proposed by Plazas and Peña [14] in the use of the same number of dummy stimuli (21) as other class members instead of two, and in the presentation of the dummy stimuli in the training phase instead of the evaluation phase.

2.3. Train Baseline Relations

Adapting the experimental trials into a data format suitable for the transformers agents corresponds to the pre-processing stage. The vocabulary consists of the class members stimuli (28), dummy stimuli (21), and response options (3), for a total of 52 tokens. Each trial is a sequence of five tokens: sample, first comparison, second comparison, third comparison, and option response. The first four tokens sequence serves as a context window for the model's training. For target data, the sequence of four tokens is shifted by one position from the same trial, resulting in first comparison, second comparison, third comparison, and the option response is then treated as the token to be predicted by the agent.

BERT and GPT architectures were trained separately on the six baseline trials sets from the experimental conditions, resulting on the 12 simulations. The model's configuration was left as default values from the original code as much as possible: six attention heads form the multi-attention head; six blocks (encoder for BERT or decoder for GPT) processed the data; a batch size of 64 trials was used; training ran for 5000 iterations; the learning rate was set to 3×10^{-4}; the number of embeddings in the attention head was 384; and a

dropout rate of 0.2 was applied. Code modifications were made in the context length to fit the size of the trials and to remove the train-validation split as the MTS procedure requires all baseline trials to be presented to the model during training.

2.4. Reflexivity Symmetry and Transitivity Evaluations

Evaluations of reflexivity, symmetry, and transitivity were conducted using specific groups of trials designed to probe each property. We employ the trial as context, akin to the prompt, comprising the sample, first comparison, second comparison, and third comparison elements. The response of the algorithm is a single token. Given the prompt (sample, comparison1, comparison2, comparison3), the expected response is one of the three option tokens (O_1, O_2, O_3). A score of 1 is assigned if the response matches the expected option, and 0 otherwise. Performance is measured by calculating the ratio of correctly selected comparisons to the total number of trials.

Despite the theoretical capability of transformers like GPT or BERT to generate any token from their vocabulary, training specifically tailors their outputs to anticipate selections from a predefined set of three options. In this study, a hallucination is defined as any instance wherein the model's response deviates from the expected selection of a pre-defined option, instead generating a stimulus from the presented class members or dummy stimuli. To analyze this phenomenon, we calculated two hallucination rates: (a) total hallucination rate, which represents the proportion of hallucinations relative to all model responses, and (b) hallucination failure rate, which represents the proportion of hallucinations specifically within incorrect model responses.

3. Results

Both BERT and GPT achieved correct selection above 98% in baseline relations across all simulations, exceeding the 90% mastery criterion for train phase across all 12 simulations, as shown in Table 1. Only under the specific condition of LS training with select–reject relations did both BERT and GPT perform above mastery criterion, with BERT achieving a slightly higher correct selection ratio. In all other conditions, neither model displayed clear evidence of equivalence class formation. Evidence of class formation was lacking in MTO or OTM train structures and when the relation type train was select-only.

Table 1. Correct selection performance of agents across train structure and relation type.

Simulation	Transformer	Train Structure	Relation Type	Baseline	Reflexivity	Symmetry	Transitivity
1	GPT	LS	S and R	0.999	0.919	0.941	0.902
2	BERT	LS	S and R	0.997	0.992	0.992	0.989
3	GPT	OTM	S and R	0.994	0.183	0.146	0.177
4	BERT	OTM	S and R	0.999	0.469	0.172	0.392
5	GPT	MTO	S and R	0.998	0.315	0.278	0.296
6	BERT	MTO	S and R	0.985	0.347	0.257	0.294
7	GPT	LS	S only	1.000	0.208	0.210	0.215
8	BERT	LS	S only	0.999	0.299	0.302	0.306
9	GPT	OTM	S only	0.999	0.068	0.016	0.075
10	BERT	OTM	S only	0.999	0.266	0.102	0.310
11	GPT	MTO	S only	1.000	0.288	0.269	0.217
12	BERT	MTO	S only	0.999	0.329	0.249	0.246

A mastery criterion is employed in matching to sample experiments to analyze performance. Arntzen [8] suggests a mastery criterion of at least 90%, in line with the work of Green and Saunders [5] of a stringent criterion to ensure accurate evidence of stimulus class formation. For a more detailed analysis, the failure range from zero to 90% correct selection ratio was subdivided into three bands of performance. The superior band limit was set above 70% as a lower mastery criterion, to evaluate close to top but insufficient performance. Another criterion, the random limit, is based on statistical character. For instance, if a random agent chooses one of three options randomly, it forms a binomial distribution with a one-third probability of a correct response. In our experiment, every

pair, namely E3-B3, is evaluated in at least 1260 trials (as a product of the permutation of 21 stimuli and three possible positions of the target comparison). Therefore, the 99.9% of possible responses of the random agent fall below 0.374, indicating that a random system typically scores below 37.46% correct selection. This delineates a differentiation between failing close to the mastery criterion and performing above a random limit, but below the soft mastery criterion.

In our simulation, BERT demonstrated reflexivity and transitivity above random level but below the lower mastery criterion in OTM with select–reject relations, albeit too low. Although failing to surpass the mastery criteria, these observations suggest attempts toward a response pattern. None of the agents passed the mastery criteria in reflexivity, transitivity, and symmetry tests across all train structures in select-only relations.

Additionally, in MTO scenarios, both GPT and BERT exhibited reflexivity of the comparison node, with BERT slightly outperforming GPT, though scores remained below the mastery level but above the soft mastery level. The performance of all 12 simulations by pair can be found in Appendix B.

The observed patterns in hallucinations offer information complimentary to the performance metric. Hallucination rate values can be found in Appendix C. These rates provide insights into the effects of TS and the presence or absence of select–reject relations on the occurrence and characteristics of hallucinations. A high total hallucination rate for GPT in tasks with select-only relations suggests a greater tendency to deviate from pre-defined options compared to BERT. Conversely, BERT's higher total hallucination rate specifically in the OTM training structure with select–reject relations suggests a potential interaction between the training structure and the presence of reject relations in influencing hallucinations. GPT's hallucination failure rate in OTM with select-only relations indicates that a significant portion of its incorrect responses were hallucinations. Conversely, BERT's substantial increase in hallucination failure rate within OTM with select–reject relations suggests that the presence of both relation types may have introduced challenges that led to more frequent incorrect responses being hallucinations.

4. Discussion

Our investigation into SE in TBM revealed no conclusive evidence of true class formation despite success in one specific condition. Both models achieved mastery criteria on reflexivity, symmetry, and transitivity evaluations within the LS structure with select–reject relations, as seen in simulations 1 and 2. However, neither model passed any tests in OTM or MTO structures, regardless of the relation type. We conducted a detailed analysis of success responses and failures for every pair across different training structures, relation types, and transformer architectures to understand the agents' outputs. Additionally, we explored the potential influence of hallucinations on their performance, discussed the contributions of this research, and outlined potential avenues for future studies.

4.1. Select–Reject Relations

The role of select–reject relations in establishing SE on humans has been documented [10–12]. This study opted to control for select relations only, excluding reject relations from the training phase. This approach aimed to isolate the agents' understanding of relationships between stimuli during training, eliminating potential influence from reject relation in test trials. Our findings emphasize the role of both select and reject relations.

In Altered MTS procedure [14], only two stimuli are used for reject relation training. This could lead to learning a simple rule of consistently rejecting those specific stimuli. To address this concern, we implemented two control measures. First, we used an equal number of dummy stimuli as other class members. Second, dummy stimuli were used exclusively during training and not presented in subsequent tests. This approach confined any potential bias associated with simple discrimination to the training phase. The evaluation phase utilized only stimuli trained in select relations, ensuring a rigorous assessment of the agents' capabilities.

Data revealed a performance drop, below random levels, in reflexivity, symmetry, and transitivity evaluation pairs when training included only select relations, as in simulations 6 (Figure A9), 7 (Figure A10), 8 (Figure A11), 9 (Figure A12), 10 (Figure A13), 11 (Figure A14), and 12 (Figure A15). This suggests that reject relations contribute to the information used by the agents to make decisions. With four stimuli in a trial, 16 possible directed relations exist, including baseline (select), reflexivity, transitivity, symmetry, and reject relations. Excluding reject relations removes information about the identification of the comparisons associated with the incorrect response to avoid.

4.2. Training Structure

The findings suggest that the ability to act as both a sample and a comparison during training is relevant for selecting the correct comparison. In the LS condition, successful performance was observed in simulation 2 (Figure A5). In contrast, in simulation 1 (Figure A4), where stimulus A acted exclusively as a sample and stimulus G as a comparison, the GPT agent's performance on those specific pairs was slightly below the mastery criterion. Although the GPT agent in simulation 1 passed the overall mastery criterion in reflexivity symmetry and transitivity tests, it did not achieve this level for all pairs (Figure A4). Additionally, the LS structure includes a greater number of stimuli (five out of seven) functioning in both sample and comparison roles compared to OTM and MTO. While OTM presents more comparison stimuli overall and MTO utilizes more sample stimuli, neither replicates the flexibility observed in the LS condition. This dissimilarity in stimulus usage across training structures likely contributes to the observed response patterns.

Previous research suggests that MTO encompasses all discriminations necessary for forming an equivalence class [8,9]. Our findings highlight the impact of specific training conditions on ANNs. Saunders and Green [9] argue that MTO holds an advantage in this regard; however, our observations suggest that certain trials within MTO may be interpreted as simple discriminations with readily learned rules. For instance, the consistent comparison in MTO could lead to a rule of "always selecting the same comparison", potentially explaining the near-mastery criterion performance for the node stimulus in reflexivity pairs in simulations 6 (Figure A9), 11 (Figure A14), and 12 (Figure A15), and the above-random-level performance in simulation 5 (Figure A8). These observations raise questions regarding the source of performance differences across TS. Factors like nodal distance and density [9] may influence results. Additionally, the role of simultaneous discrimination [5,39] seems less prominent in this context. The observed discrepancies in performance between TBM likely stem from other processes, requiring further exploration.

4.3. Computational Agents Architecture

The architectural differences between GPT and BERT play a role in their performance on SE tasks. Unlike GPT, BERT lacks a mask in its self-attention head, enabling it to process information about both preceding and subsequent tokens bidirectionally [25,26,28]. This is essential for analyzing how each stimulus communicates and forms its network of connections and may seem relevant during training when the response token is the final sequence element. It is important to note that, in our experiments, BERT does not gain additional context beyond that response token. Both models receive the same information regarding the trial tokens before the response. Therefore, the difference lies not in accessing the response itself, as it is never part of the context. This communication within the attention heads occurs solely among the class member stimuli during the trial.

BERT's bidirectional processing allows it to analyse information flowing both forwards and backwards. This additional information may contribute to the development of new rules, potentially explaining its superior performance in the LS condition with select–reject relations. The combination of this structure with reject relations may be sufficient for BERT to respond correctly. In contrast, GPT utilizes a masked multi-head self-attention mechanism, permitting only unidirectional information flow within its communication and token attention functions.

The concept of feature representation within attention heads offers a potential explanation for the models' rule formation. These representations may allow them to develop internal decision-making rules for selecting the correct answer. These rules may not necessarily reflect SE. Instead, they could be a combination of discrimination and learned reject relations. This could explain the observed performance discrepancies across training structures. LS structure, with its flexibility in stimulus usage, may be better suited for learning these combined rules compared to MTO and OTM, which offer less variation in stimulus presentation.

The analysis of hallucinations rates presented in Table A4 revealed a potential link with performance. Cases wherein models with high hallucination rates also displayed low accuracy (e.g., GPT in OTM with select-only relations) suggest that hallucinations may be indicative of the model struggling with the task. Conversely, situations wherein models exhibited a high hallucination failure rate coinciding with moderate accuracy levels (e.g., BERT in OTM with select–reject relations) suggest that the model may be attempting responses but encountering difficulties, resulting in either correct answers or significant errors as hallucinations. One potential explanation for this phenomenon lies in the limitations of using SE tasks for evaluating transformer models. Unlike controlled psychological experiments wherein responses are strictly limited to pre-defined options, the current adaptation of SE tasks allow models to exploit the presence of class members as potential response tokens. This could lead to incorrect selections when information from other tokens (e.g., lack of reject relations) is missing, potentially manifesting as hallucinations. BERT displayed a tendency to generate transitivity responses above chance levels but below mastery criteria specifically in OTM with select–reject relations. This suggests that BERT may be attempting to learn the underlying relations but encountering difficulties within the OTM training structure, potentially leading to hallucinations as it attempts alternative response strategies.

4.4. General Discussion

This study investigated the performance of TBMs in SE tasks. Overall, BERT demonstrated a slight edge over GPT, but both models respond similarly to changes in experimental conditions. Agents performance in LS suggests that features represented in the attention heads may facilitate rule selection in linearly ordered tasks. Both feed-forward networks [19] and TBMs like the ones tested here can perform conditional discrimination tasks, as evidenced by success rates in baseline training. They seem to lack the mechanisms necessary for responding based on established equivalence classes. Correct responses from TBMs can be explained as a combination of discrimination process and reject relations and may not indicate equivalence class formation.

We found similarities in general performance and small differences when comparing TBMs with other models. While feed-forward networks in Carrillo and Betancort's study [19] demonstrated some success in specific conditions, our TBMs exhibited superior performance, in terms of achieving mastery criteria on baseline, reflexivity, symmetry, and transitivity tests under select–reject conditions in LS. Additionally, our inclusion of the select-only condition, not explored in Carrillo and Betancort's study, provides a more comprehensive assessment of model capabilities. A key difference between TBMs and the EPS/E-EPS models [20,21] lies in their ability to replicate human performance in stimulus equivalence tasks. While EPS and E-EPS have demonstrated success in reproducing the outcomes of prominent studies, our TBMs encountered challenges in OTM and MTO, suggesting limitations in capturing the full complexity of human equivalence class formation. EPS and E-EPS models have primarily focused on the influence of different training structures on equivalence class formation. However, the role of relation type has not been extensively explored in these models. In contrast, our study investigated the impact of both training structure and relation type on TBM performance, providing a more comprehensive understanding of the factors influencing equivalence class formation.

4.4.1. Lack of Evidence of an Actual Equivalence Response

While BERT and GPT achieved success in LS with select–reject relations, their limitations in other training configurations do not support SE achievement. Their inability in MTO and OTM structures, coupled with their failure across all structures when trained with select-only relations, raises questions about their capacity for true SE. These observations challenge the claims that transformers are capable of abstract reasoning and symbolic manipulation [32]. Their success in LS with select–reject relations may be attributed to a form of pattern recognition and memorization specific to this structure's relational pathways [33]. The presence of both correct and reject relations potentially guides the models toward the desired response by offering a more constrained choice environment. This interpretation aligns with their struggles in MTO, OTM, and select-only scenarios. These structures lack the relational path and explicit rejection cues present in LS with select–reject. Without these features, transformers fail both in the test of equivalence and to generalize their learnings to new relational contexts.

Humans, on the other hand, can achieve stimulus class formation regardless of training structure and relation types [1,5,9]. This suggests a more flexible and generalizable understanding of equivalence compared to current FLMs. These findings highlight the gap between FLMs and human SE capabilities. While FLMs show some promise in specific scenarios, their limitations suggest a reliance on pattern recognition rather than abstract reasoning.

4.4.2. Contributions

This research introduces SE as a novel tool for probing abstraction and symbolic manipulation capabilities of TBMs beyond memorization and pattern recognition. By analyzing TBM performance in SE tasks, which require forming abstract relationships between stimuli, we gain insights into their capacity. This approach indicates the viability of SE as an explainability technique [40] and an FLM benchmark.

Traditional benchmarks often assess FLMs on specific tasks, which heavily rely on pre-acquired knowledge [41,42]. Commonly evaluated tasks are factual knowledge [43], commonsense inference [44] and reasoning [45], model's ability to apply knowledge across various domains [46], tendency to generate falsehoods [47], or mathematical reasoning [48]. The focus on SE with minimal training data allows researchers to isolate and evaluate learning mechanisms in FLMs as an intrinsic feature of their design. Traditional benchmarks often rely on massive datasets, potentially obscuring the underlying learning processes. By utilizing SE with minimal data requirements, we gain a deeper understanding of how TBMs learn and generalize knowledge in resource-constrained environments. Analyzing model performance in tasks requiring reasoning and abstraction with SE as a benchmark contributes to the development of more transparent and trustworthy systems. SE offers a valuable benchmark by focusing on core learning abilities independent of pre-training or fine-tuning, evaluating a model's ability to learn and apply generalizable rules.

This study demonstrates the value of exploring SE tasks to gain insights into FLMs' capabilities beyond narrowly defined tasks. This approach aligns with the goals of the Abstraction and Reasoning Corpus (ARC) dataset [49], which emphasizes assessing general intelligence through abstract reasoning and efficient skill acquisition. Experiments with SE tasks could provide evidence for the models' ability to generalize knowledge, transfer learning, and symbolic representation and manipulation. Both SE and ARC frameworks assess processes beyond memorizing specific training examples. SE tasks offer a controlled environment to probe these processes, allowing researchers to evaluate FLMs' ability in abstract reasoning. Other proposals of measuring abstract reasoning are based on Raven's Progressive Matrices [50]. SE has been used in humans also as an assessment tool for intelligence [51].

Studying SE in DL models offers a comparative approach to understanding human cognition. This work highlights limitations of TBMs in generalizing knowledge and forming true equivalence classes. These findings inform potential areas for further human SE research, like the role of working memory or attention in overcoming these limita-

tions observed in TBMs. Furthermore, standardized SE tasks used with FLMs can be adapted to study developmental differences in human SE learning or validate existing psychological models.

The performance of both models in the LS training structure with select–reject relations warrants exploration through the lens of CoT prompting. BERT's performance, exceeding mastery level on all stimulus pairs in LS with select–reject relations, suggests an ability to process the sequence of stimuli and identify underlying relationships. This aligns with the step-by-step approach in CoT, wherein the model goes through connections between elements to reach a conclusion. GPT's success with symmetry and transitivity pairs wherein stimuli acted as both sample and comparison during baseline relation training further supports this notion.

GPT's challenges with specific transitivity pairs when the order of stimuli is reversed to the role in baseline relations, as in A as comparison and G as sample, on those stimuli which were not used as both sample and comparison, highlight potential limitations in its capacities. This selective difficulty suggests a barrier to generalizing learned relations and applying them in different contexts. Unlike BERT, GPT may not be consistently following structured decision rules within the LS structure. These findings align with the documented phenomenon of the reversal curse [36]. In the context of SE tasks, GPT's struggles with reversed transitivity pairs suggest difficulties generalizing the underlying equivalence relation when the stimulus order is flipped. This highlights a potential limitation in GPT's capabilities compared to BERT's success in processing the LS structure. In MTO and OTM, both BERT and GPT failed on symmetry pairs. This is also consistent with reversal learning difficulties.

The ability of GPT and BERT to respond correctly on transitivity, reflexivity, and symmetry in the SE tasks with select–reject relations in the LS structure can be related to the concept of few-shot learning. These models were not explicitly trained on all possible relations within the SE task. They only received examples of baseline pairs. Despite this limited training, both models demonstrated the ability to respond to relations beyond those directly trained on. This suggests that transformer models with bidirectional processing like BERT may be capable of learning underlying relations and generalizing to new stimuli with minimal exposure.

There is a relation of the findings with Reinforcement Learning from Human Feedback (RLHF) for fine-tuning language models. In our simulations, models exposed to both select–reject relations during LS training learned to respond correctly. RLHF uses human feedback as a learning signal [52], with positive feedback reinforcing select relations between the model's output and the desired response, while negative feedback acts as a reject relation. However, RLHF faces challenges due to the difficulty of defining all incorrect responses. While FLMs can learn to reject demonstrably wrong outputs, exhaustively mapping all possible incorrect answer tokens is impractical. This limitation creates the potential for hallucinations, wherein the model selects seemingly plausible but ultimately incorrect responses during generation. In theory, models with the ability to form equivalent classes should not be so dependent on RLHF. SE capabilities could direct them to the correct answer without further feedback needed.

4.4.3. Limitations and Further Research

Limitations were considered and posed as future research possibilities. We employed a restricted set of variables, such as training structures and relation types. A broader range of conditions to gain a more comprehensive understanding of the factors influencing SE can be explored. This could lead to important contributions to the understanding of SE in both humans and DL models. Additionally, investigating the impact of pre-training data or fine-tuning would be also informative.

This research employed minimal code implementations of BERT and GPT with a size of 10 million parameters using local hardware, while state-of-the-art FLMs often possess billions of parameters. This limited model may restrict the complexity of learned

relationships and hinder the generalizability of the findings to more powerful FLMs. Future research should prioritize upscaling the models using more powerful hardware and lager training data volumes. Investigating a broader range of FLM architectures may provide a more comprehensive understanding of how SE performance varies across model types. The focus on limited training data strengthens the investigation of learning abilities, but it also limits the generalizability of the findings. Future investigations can also explore the effect of pre-training or fine-tuning on the performance of FLMs in SE experiments.

Despite exploring hallucinations, the internal decision-making processes of TBMs within SE tasks remain largely opaque. Integrating SE to explainability techniques [40] would offer insights into their decision-making mechanisms. Alternative training paradigms that better simulate the controlled response options present in psychological experiments to control for the influence of reject relations on hallucinations can be implemented. Exploring training techniques specifically designed to mitigate the reversal curse in unidirectional models like GPT could involve incorporating training paradigms that encourage backward reasoning or utilize auxiliary tasks that promote bidirectional information processing within these models.

5. Conclusions

This investigation explored the capabilities of TBMs in an SE experiment. While both BERT and GPT achieved success in the LS structure with select–reject relations, their performance limitations in other configurations suggest the formation of decision rules based on a combination of discrimination and select–reject relations. The inability to respond correctly in OTM and MTO structures, coupled with failures across all structures when trained with select-only relations, differs from human performance and falls short of conclusive equivalence class formation. Our findings suggest that the flexibility to utilize stimuli as both samples and comparisons during training appears to be a critical factor for agents' response rules. The reliance of the models on reject relations and comparison stimuli information affects them in their performance and in the occurrence of hallucinations. Response patterns of TBMs on SE experiments align with the concept of few-shot learning and the reversal curse.

We highlight the potential of SE as a explainability technique and a benchmark for evaluating learning and reasoning abilities in FLMs, independent of pre-training on massive datasets or fine-tuning. SE tasks require applying learned relationships to novel stimuli, demonstrating the model's ability to generalize knowledge beyond specific training examples. Success in SE tasks could suggests that the model can transfer learning from the training context to new situations and manipulate symbols. Compared to traditional benchmarks, SE focuses on core mechanisms and minimal training data requirements. SE resonates with the ARC goals by assessing abstract reasoning and transfer learning capabilities in TBM. Both frameworks offer insights into processes beyond memorization. SE tasks may involve the internal representation of stimuli and manipulation of those representations to identify equivalence classes. This could shed light on the models' ability to form and manipulate symbolic representations.

Author Contributions: Supervision, M.B.; writing, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The simulations' code can be found in at https://github.com/Yagwar/stim_eq/tree/master/SE_GPT (accessed on 5 August 2024). There is a trials creator script and a notebook that executes train and evaluation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TBM	Transformer-Based Model
SE	Stimulus Equivalence
MTS	Matching-To-Sample
ANNs	Artificial Neural Networks
DL	Deep Learning
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pretrained Transformer
FLM	Foundational Language Models
CoT	Chain-of-Thought
TS	Train Structure
LS	Linear Series
MTO	Many-To-One
OTM	One-To-Many
RLHF	Reinforcement Learning from Human Feedback

Appendix A. Train Structures

Tables A1–A3 presented in this appendix outline the specific stimulus pairs included in each training condition (LS, OTM and MTO). Each table provides a matrix where rows represent sample stimuli and columns represent comparison stimuli. Cell entries indicate the pair subset to which each stimulus combination belongs based on the training structure. The accompanying figures visually depict the relationships between stimuli within each equivalence class. The directed graphs illustrate baseline, symmetry, transitivity, and reflexivity pairs.

Appendix A.1. Linear Series

Table A1. LS pairs and their subset type.

Sample	Comparison						
	A	**B**	**C**	**D**	**E**	**F**	**G**
A	*A-A* *Reflexivity*	**A-B** **Baseline**	A-C Transitivity	A-D Transitivity	A-E Transitivity	A-F Transitivity	A-G Transitivity
B	***B-A*** ***Symmetry***	*B-B* *Reflexivity*	**B-C** **Baseline**	B-D Transitivity	B-E Transitivity	B-F Transitivity	B-G Transitivity
C	C-A Transitivity	***C-B*** ***Symmetry***	*C-C* *Reflexivity*	**C-D** **Baseline**	C-E Transitivity	C-F Transitivity	C-G Transitivity
D	D-A Transitivity	D-B Transitivity	***D-C*** ***Symmetry***	*D-D* *Reflexivity*	**D-E** **Baseline**	D-F Transitivity	D-G Transitivity
E	E-A Transitivity	E-B Transitivity	E-C Transitivity	***E-D*** ***Symmetry***	*E-E* *Reflexivity*	**E-F** **Baseline**	E-G Transitivity
F	F-A Transitivity	F-B Transitivity	F-C Transitivity	F-D Transitivity	***F-E*** ***Symmetry***	*F-F* *Reflexivity*	**F-G** **Baseline**
G	G-A Transitivity	G-B Transitivity	G-C Transitivity	G-D Transitivity	G-E Transitivity	***G-F*** ***Symmetry***	*G-G* *Reflexivity*

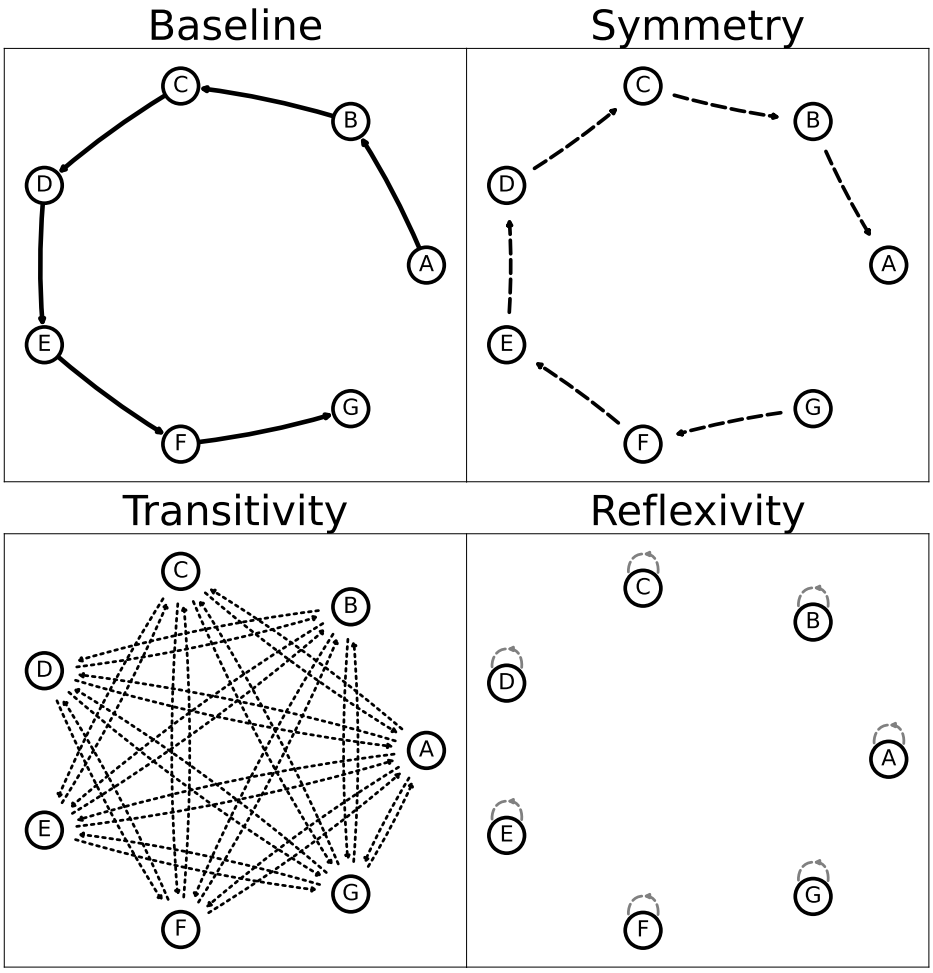

Figure A1. Linear series train structure pairs by subset. Top left panel shows Baseline relations in black solid arrows. Symmetry in black dashed arrows on top right panel. Bottom left shows transitivity pairs in black dotted arrows. Bottom right shows reflexivity in dashed grey arrows.

Appendix A.2. One-to-Many Train Structure

Table A2. OTM pairs and their subset type.

Sample	Comparison						
	A	**B**	**C**	**D**	**E**	**F**	**G**
A	*A-A* *Reflexivity*	**A-B** **Baseline**	**A-C** **Baseline**	**A-D** **Baseline**	**A-E** **Baseline**	**A-F** **Baseline**	**A-G** **Baseline**
B	***B-A*** ***Symmetry***	*B-B* *Reflexivity*	B-C Transitivity	B-D Transitivity	B-E Transitivity	B-F Transitivity	B-G Transitivity
C	***C-A*** ***Symmetry***	C-B Transitivity	*C-C* *Reflexivity*	C-D Transitivity	C-E Transitivity	C-F Transitivity	C-G Transitivity
D	***D-A*** ***Symmetry***	D-B Transitivity	D-C Transitivity	*D-D* *Reflexivity*	D-E Transitivity	D-F Transitivity	D-G Transitivity
E	***E-A*** ***Symmetry***	E-B Transitivity	E-C Transitivity	E-D Transitivity	*E-E* *Reflexivity*	E-F Transitivity	E-G Transitivity
F	***F-A*** ***Symmetry***	F-B Transitivity	F-C Transitivity	F-D Transitivity	F-E Transitivity	*F-F* *Reflexivity*	F-G Transitivity
G	***G-A*** ***Symmetry***	G-B Transitivity	G-C Transitivity	G-D Transitivity	G-E Transitivity	G-F Transitivity	*G-G* *Reflexivity*

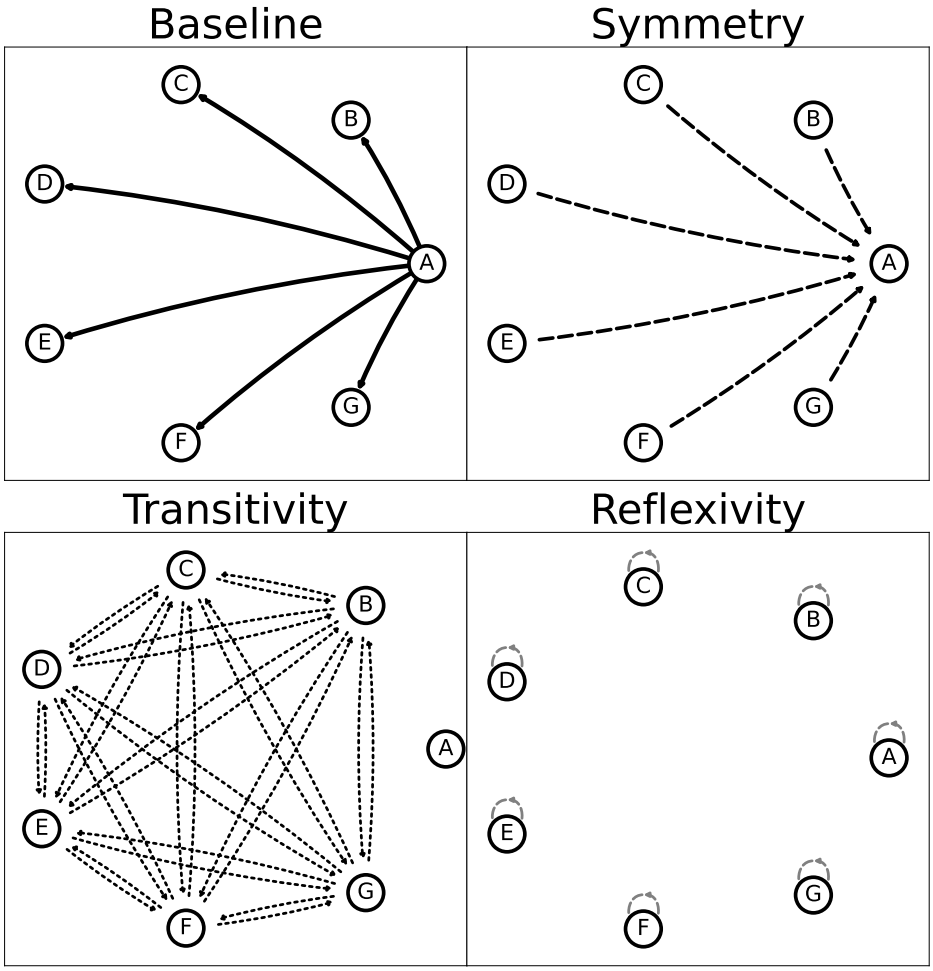

Figure A2. One-to-many. Top left panel shows Baseline relations in black solid arrows. Symmetry in black dashed arrows on top right panel. Bottom left panel shows transitivity pairs in black dotted arrows. Bottom right panel shows reflexivity in dashed grey arrows.

Appendix A.3. Many-to-One Train Structure

Table A3. MTO pairs and their subset type.

Sample	Comparison						
	A	**B**	**C**	**D**	**E**	**F**	**G**
A	*A-A* *Reflexivity*	***A-B*** ***Symmetry***	***A-C*** ***Symmetry***	***A-D*** ***Symmetry***	***A-E*** ***Symmetry***	***A-F*** ***Symmetry***	***A-G*** ***Symmetry***
B	**B-A** **Baseline**	*B-B* *Reflexivity*	B-C Transitivity	B-D Transitivity	B-E Transitivity	B-F Transitivity	B-G Transitivity
C	**C-A** **Baseline**	C-B Transitivity	*C-C* *Reflexivity*	C-D Transitivity	C-E Transitivity	C-F Transitivity	C-G Transitivity
D	**D-A** **Baseline**	D-B Transitivity	D-C Transitivity	*D-D* *Reflexivity*	D-E Transitivity	D-F Transitivity	D-G Transitivity
E	**E-A** **Baseline**	E-B Transitivity	E-C Transitivity	E-D Transitivity	*E-E* *Reflexivity*	E-F Transitivity	E-G Transitivity
F	**F-A** **Baseline**	F-B Transitivity	F-C Transitivity	F-D Transitivity	F-E Transitivity	*F-F* *Reflexivity*	F-G Transitivity
G	**G-A** **Baseline**	G-B Transitivity	G-C Transitivity	G-D Transitivity	G-E Transitivity	G-F Transitivity	*G-G* *Reflexivity*

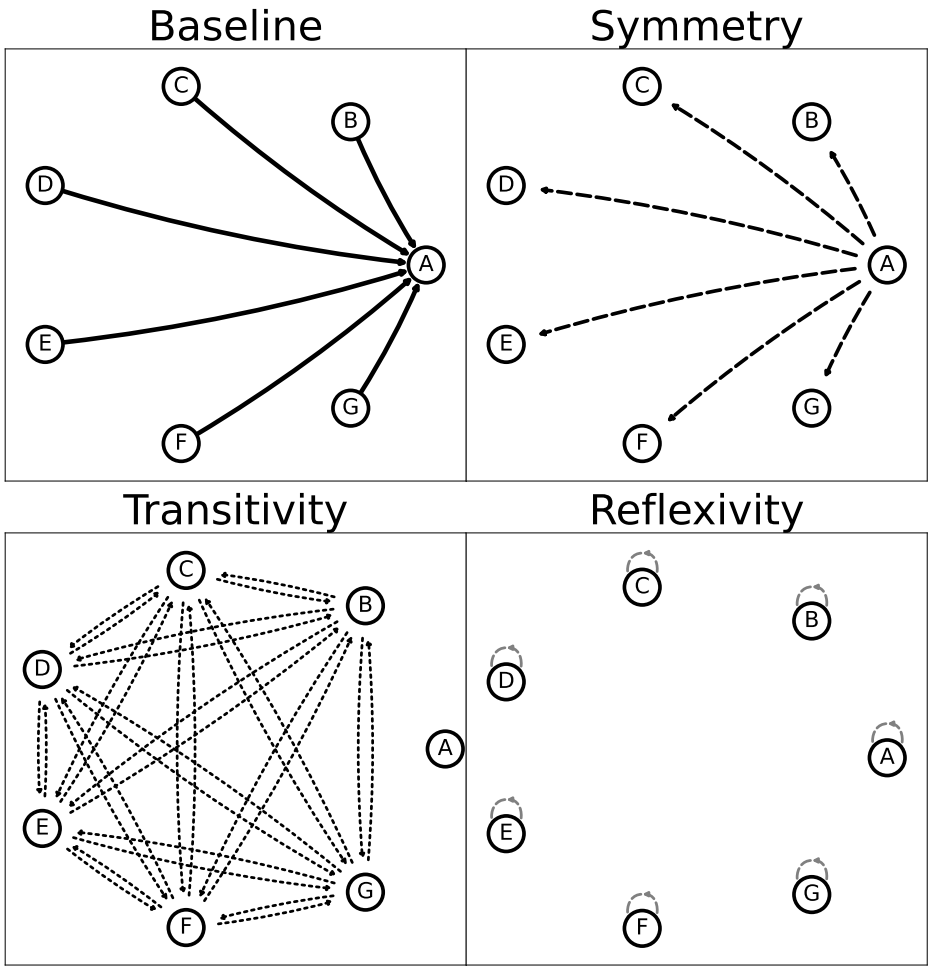

Figure A3. Many-to-one. Top left panel shows Baseline relations in black solid arrows. Symmetry in black dashed arrows on top right panel. Bottom left panel shows transitivity pairs in black dotted arrows. Bottom right panel shows reflexivity in dashed grey arrows.

Appendix B. Simulations Pairs Performance

Figures A4–A15 presents the performance metrics and equivalence class formation for the 12 simulated experiments. Each figure is divided into two panels. The left panel displays a heatmap matrix representing the correct selection ratio for each stimulus pair, color-coded to indicate performance levels. Blue indicates mastery-level performance (0.9–1.0), purple denotes below-mastery performance (0.7–0.9), orange represents above-random performance (0.37–0.7), and red signifies below-random performance (0–0.37). The right panel shows a directed graph illustrating the stimulus pairs based on their training structure. Solid arrows represent baseline relationships, dashed self-loops indicate reflexivity, dashed arrows represent symmetry, and dotted arrows represent transitivity. The color of the arrows corresponds to the performance level of the respective pair.

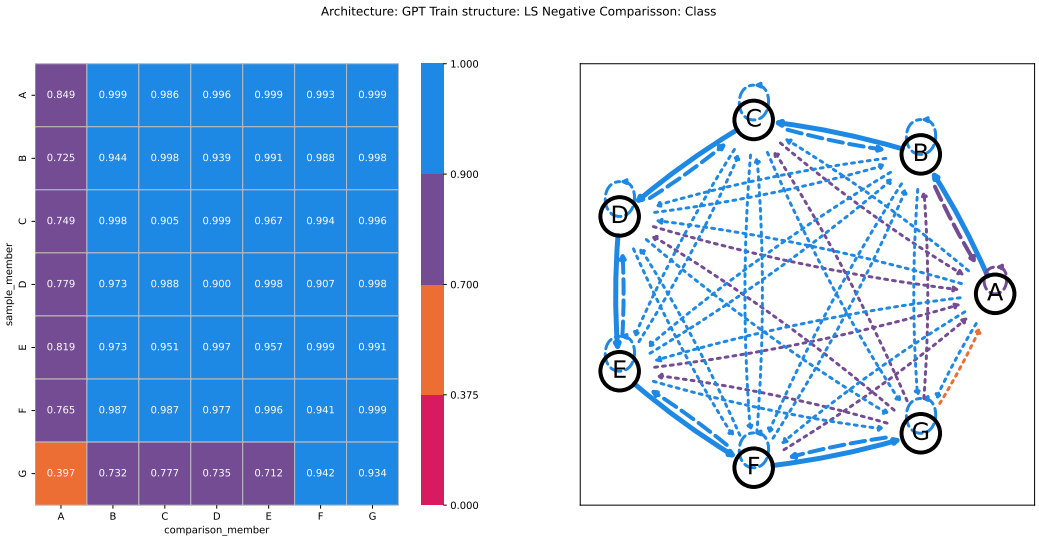

Figure A4. Simulation 1 performance metrics.

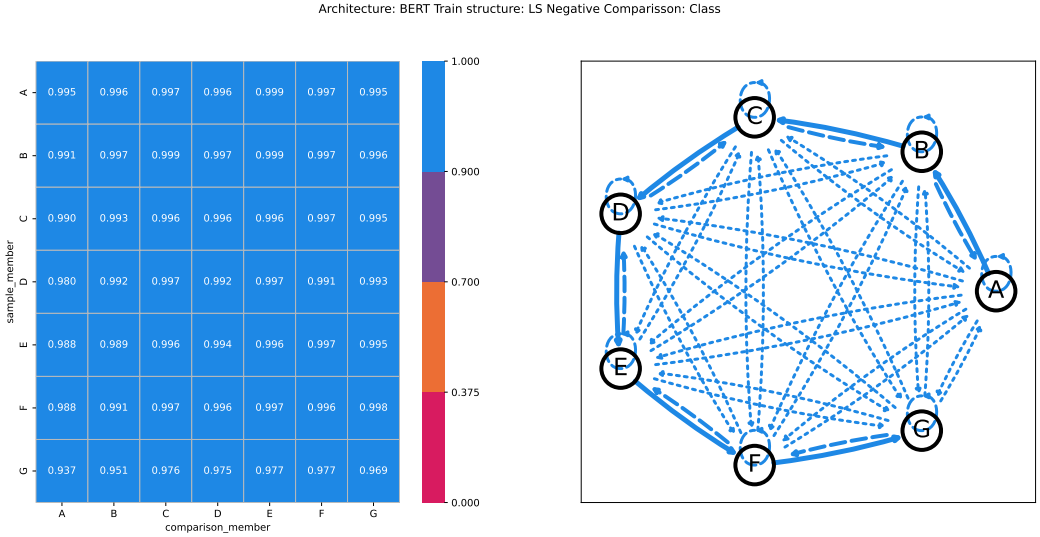

Figure A5. Simulation 2 performance metrics.

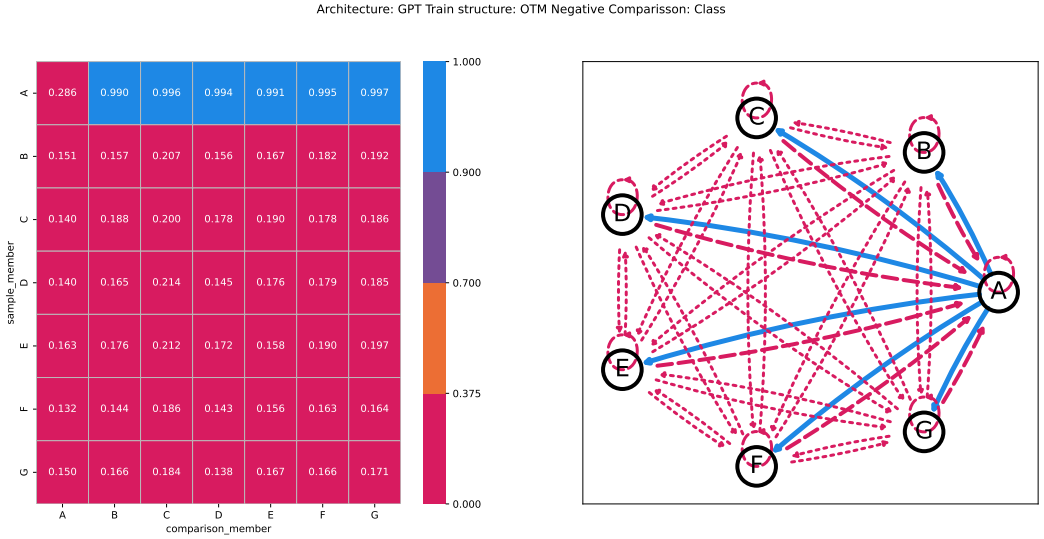

Figure A6. Simulation 3 performance metrics.

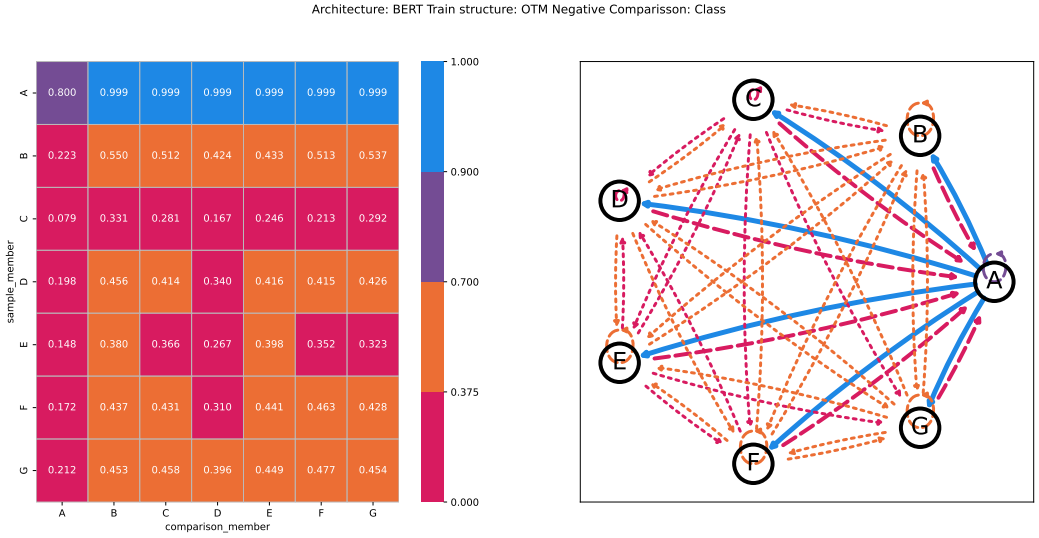

Figure A7. Simulation 4 performance metrics.

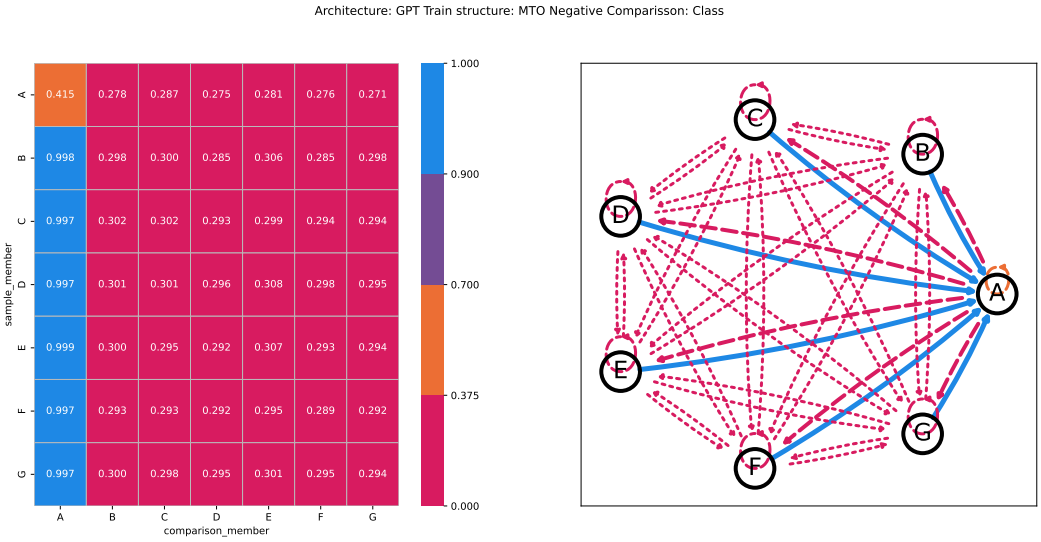

Figure A8. Simulation 5 performance metrics.

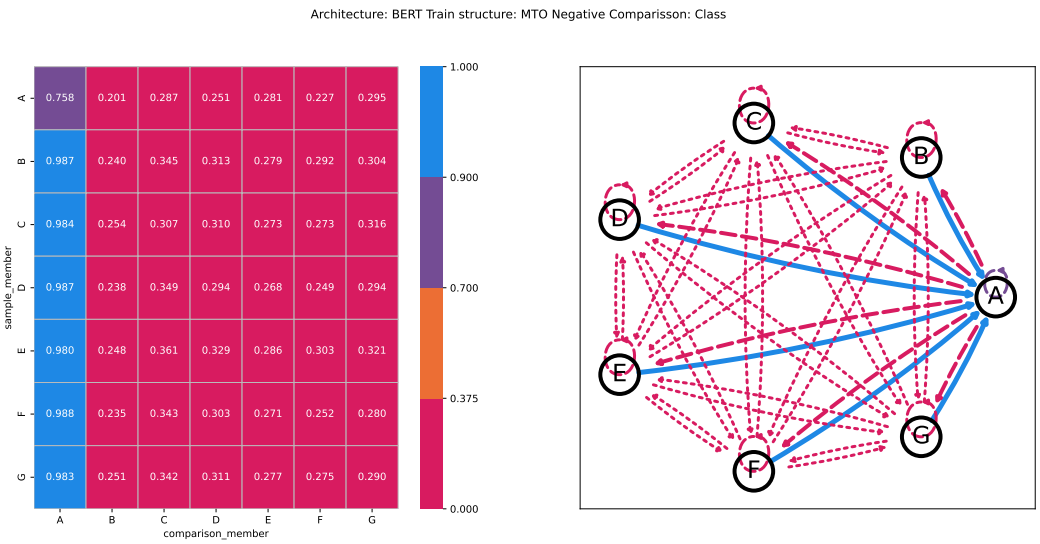

Figure A9. Simulation 6 performance metrics.

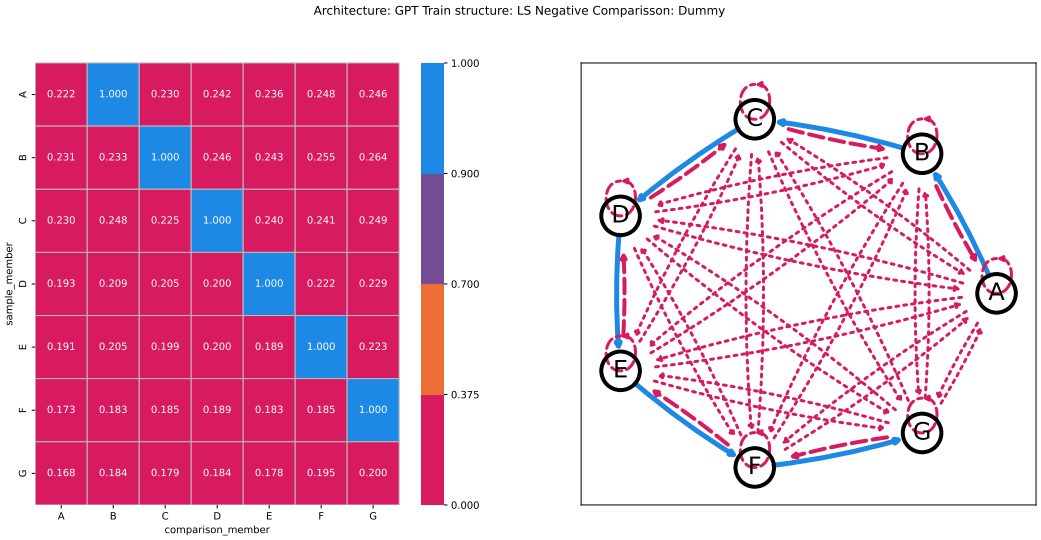

Figure A10. Simulation 7 performance metrics.

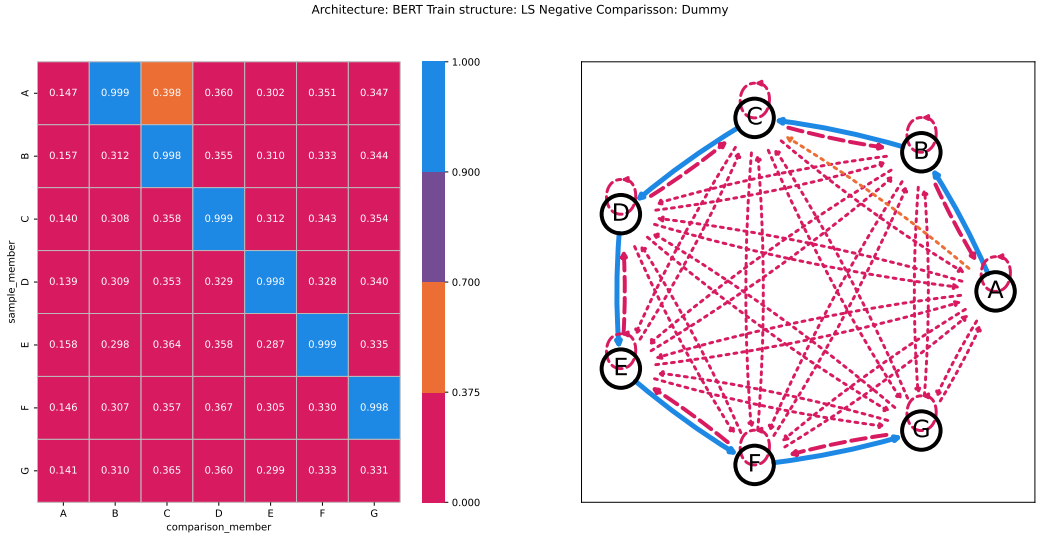

Figure A11. Simulation 8 performance metrics.

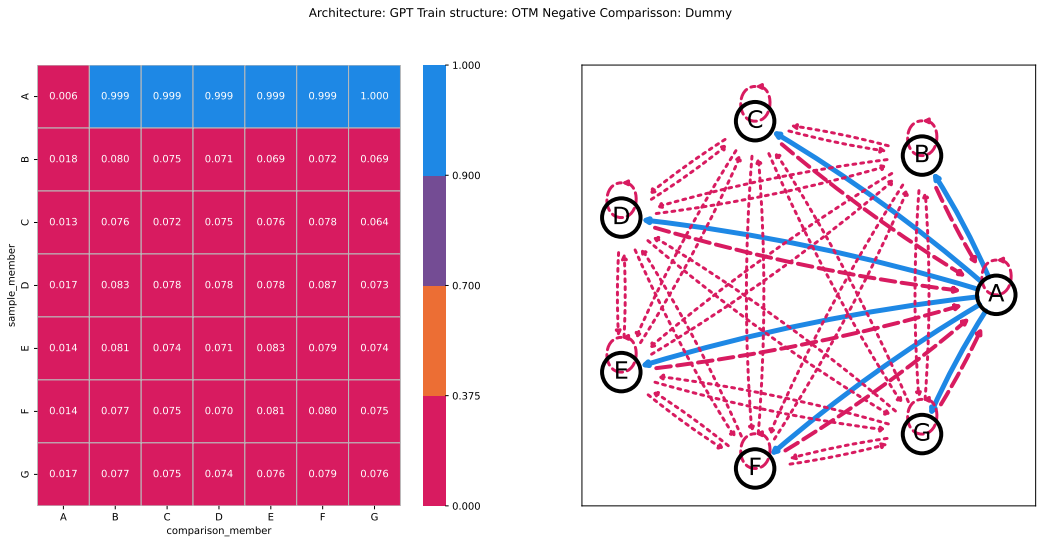

Figure A12. Simulation 9 performance metrics.

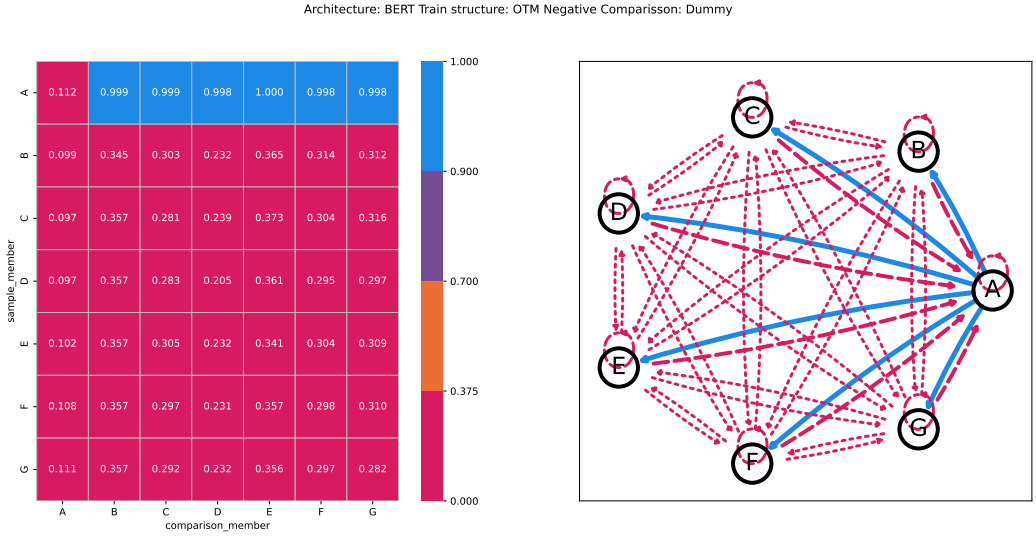

Figure A13. Simulation 10 performance metrics.

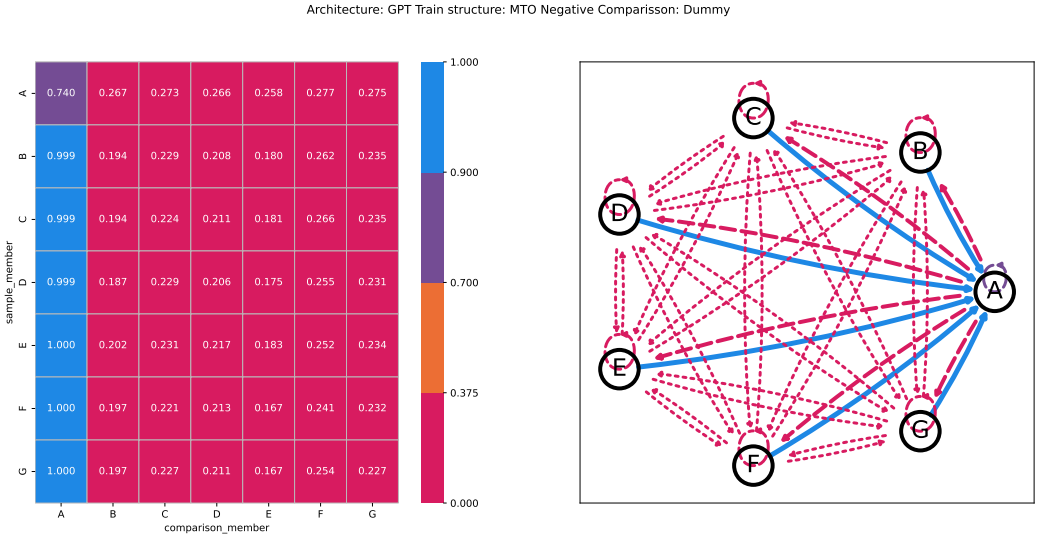

Figure A14. Simulation 11 performance metrics.

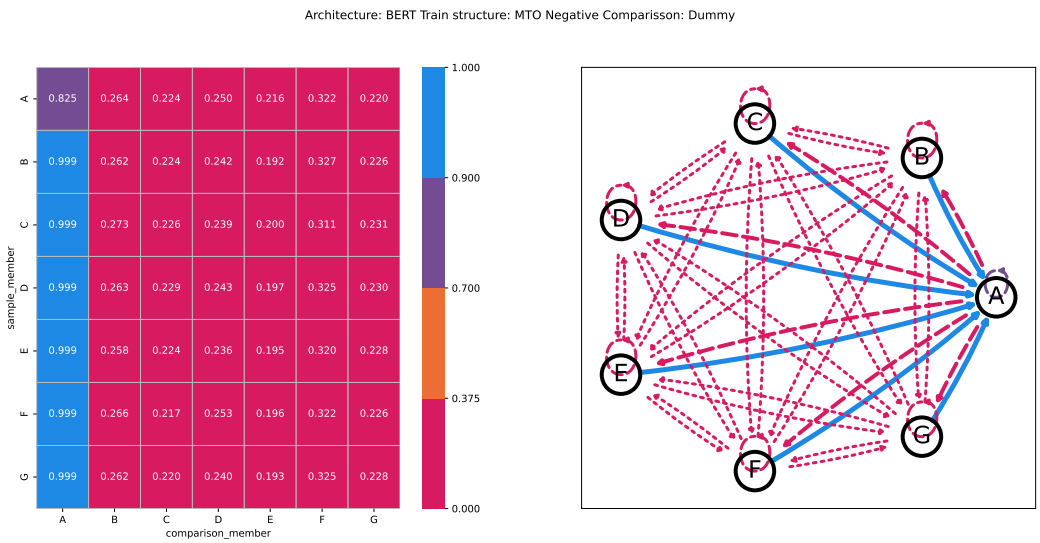

Figure A15. Simulation 12 performance metrics.

Appendix C. Hallucination Analysis

Table A4. Hallucination rates.

Simulation	Transformer	Train Structure	Relation Type	Total Hallucination Rate				Hallucination Fail Rate			
				Baseline	Reflexivity	Symmetry	Transitivity	Baseline	Reflexivity	Symmetry	Transitivity
1	**GPT**	**LS**	**S and R**	0.000	0.004	0.001	0.001	0.188	0.045	0.012	0.009
2	**BERT**	**LS**	**S and R**	0.000	0.000	0.000	0.000	0.023	0.013	0.000	0.005
3	**GPT**	**OTM**	**S and R**	0.001	0.016	0.011	0.014	0.112	0.020	0.013	0.016
4	**BERT**	**OTM**	**S and R**	0.000	0.275	0.424	0.321	0.240	0.518	0.512	0.527
5	**GPT**	**MTO**	**S and R**	0.000	0.030	0.001	0.032	0.081	0.043	0.002	0.046
6	**BERT**	**MTO**	**S and R**	0.001	0.044	0.030	0.021	0.037	0.067	0.041	0.030
7	**GPT**	**LS**	**S only**	0.000	0.431	0.368	0.351	0.571	0.544	0.466	0.448
8	**BERT**	**LS**	**S only**	0.001	0.103	0.082	0.071	0.488	0.148	0.118	0.102
9	**GPT**	**OTM**	**S only**	0.001	0.794	0.539	0.832	1.000	0.852	0.547	0.900
10	**BERT**	**OTM**	**S only**	0.001	0.183	0.235	0.140	0.436	0.249	0.261	0.203
11	**GPT**	**MTO**	**S only**	0.000	0.125	0.002	0.147	0.417	0.175	0.002	0.188
12	**BERT**	**MTO**	**S only**	0.000	0.000	0.001	0.000	0.276	0.001	0.001	0.000

References

1. Sidman, M. Equivalence relations and the reinforcement contingency. *J. Exp. Anal. Behav.* **2000**, *74*, 127–146. [CrossRef] [PubMed]
2. Sidman, M.; Tailby, W. Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *J. Exp. Anal. Behav.* **1982**, *37*, 5–22. [CrossRef] [PubMed]
3. Alonso-Alvarez, B. The Problem of Class Breakdown in Sidman's (1994, 2000) Theory about the Origin of Stimulus Equivalence. *Perspect. Behav. Sci.* **2023**, *46*, 217–235. [CrossRef] [PubMed]
4. Sidman, M. What Is Interesting about Equivalence Relations and Behavior? *Perspect. Behav. Sci.* **2018**, *41*, 33–43. [CrossRef]
5. Green, G.; Saunders, R.R. Stimulus Equivalence. In *Handbook of Research Methods in Human Operant Behavior*; Lattal, K.A., Perone, M., Eds.; Springer: Boston, MA, USA, 1998; pp. 229–262. [CrossRef]
6. Tovar, Á.E.; Torres-Chávez, Á.; Mofrad, A.A.; Arntzen, E. Computational models of stimulus equivalence: An intersection for the study of symbolic behavior. *J. Exp. Anal. Behav.* **2023**, *119*, 407–425. [CrossRef] [PubMed]
7. Sidman, M. Equivalence relations and behavior: An introductory tutorial. *Anal. Verbal Behav.* **2009**, *25*, 5–17. [CrossRef] [PubMed]
8. Arntzen, E. Training and testing parameters in formation of stimulus equivalence: Methodological issues. *Eur. J. Behav. Anal.* **2012**, *13*, 123–135. [CrossRef]
9. Saunders, R.R.; Green, G. A Discrimination Analysis of Training-Structure Effects on Stimulus Equivalence Outcomes. *J. Exp. Anal. Behav.* **1999**, *72*, 117–137. [CrossRef] [PubMed]
10. Carrigan, P.F., Jr.; Sidman, M. Conditional Discrimination and Equivalence Relations: A Theoretical Analysis of Control by Negative Stimuli. *J. Exp. Anal. Behav.* **1992**, *58*, 183–204. [CrossRef]
11. Johnson, C.; Sidman, M. Conditional Discrimination and Equivalence Relations: Control by Negative Stimuli. *J. Exp. Anal. Behav.* **1993**, *59*, 333–347. [CrossRef]
12. Plazas, E.A. Formation of Stimulus Equivalence Relations by Exclusion: Evidence using the Blank Comparison Stimulus Procedure. *Psychol. Rec.* **2021**, *71*, 1–15. [CrossRef]
13. Hinojo Abujas, Z.; Pérez Fernández, V.; García García, A. The formation of equivalence classes in adults without training in negative relations between members of different classes. *Int. J. Psychol. Psychol. Ther.* **2017**, *17*, 107–118.
14. Plazas, E.A.; Peña, T.E. Effects of Procedural Variations in the Training of Negative Relations for the Emergence of Equivalence Relations. *Psychol. Rec.* **2016**, *66*, 109–125. [CrossRef]
15. Barnes, D.; Hampson, P.J. Stimulus Equivalence and Connectionism: Implications for Behavior Analysis and Cognitive Science. *Psychol. Rec.* **1993**, *43*, 617–638. [CrossRef]
16. Tovar, A.E.; Chávez, A.T. A Connectionist Model of Stimulus Class Formation with a Yes/No Procedure and Compound Stimuli. *Psychol. Rec.* **2012**, *62*, 747–762. [CrossRef]
17. Vernucio, R.R.; Debert, P. Computational Simulation of Equivalence Class Formation Using the go/no-go Procedure with Compound Stimuli. *Psychol. Rec.* **2016**, *66*, 439–449. [CrossRef] [PubMed]
18. Ninness, C.; Ninness, S.K.; Rumph, M.; Lawson, D. The Emergence of Stimulus Relations: Human and Computer Learning. *Perspect. Behav. Sci.* **2018**, *41*, 121–154. [CrossRef] [PubMed]
19. Carrillo, A.; Betancort, M. Differences of Training Structures on Stimulus Class Formation in Computational Agents. *Multimodal Technol. Interact.* **2023**, *7*, 39. [CrossRef]
20. Mofrad, A.A.; Yazidi, A.; Hammer, H.L.; Arntzen, E. Equivalence Projective Simulation as a Framework for Modeling Formation of Stimulus Equivalence Classes. *Neural Comput.* **2020**, *32*, 912–968. [CrossRef]
21. Mofrad, A.A.; Yazidi, A.; Mofrad, S.A.; Hammer, H.L.; Arntzen, E. Enhanced Equivalence Projective Simulation: A Framework for Modeling Formation of Stimulus Equivalence Classes. *Neural Comput.* **2021**, *33*, 483–527. [CrossRef]
22. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA , 2016; p. 800.
23. Alpaydin, E. *Introduction to Machine Learning*, 4th ed.; MIT Press: Cambridge, MA, USA, 2020.
24. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805.
27. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 15 January 2024).
28. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
29. Bowen, C.; Sætre, R.; Miyao, Y. A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models. In *Findings of the Association for Computational Linguistics: Proceedings of the EACL 2024, St. Julian's, Malta, 18–22 March 2024*; Graham, Y., Purver, M., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 323–339.
30. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: Proceedings of the ACL 2023, Toronto, ON, Canada, 9–14 July 2023*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 1049–1065. [CrossRef]

31. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2201.11903.
32. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682.
33. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21), New York, NY, USA, 3–10 March 2021; pp. 610–623. [CrossRef]
34. Schaeffer, R.; Miranda, B.; Koyejo, S. Are Emergent Abilities of Large Language Models a Mirage? In *Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 55565–55581.
35. Mahowald, K.; Ivanova, A.A.; Blank, I.A.; Kanwisher, N.; Tenenbaum, J.B.; Fedorenko, E. Dissociating language and thought in large language models. *Trends Cogn. Sci.* **2024**, *28*, 517–540. [CrossRef] [PubMed]
36. Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A.C.; Korbak, T.; Evans, O. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv* **2024**, arXiv:2309.12288.
37. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.01219.
38. Karpathy, A. ng-video-lecture. 2023. Available online: https://github.com/karpathy/ng-video-lecture (accessed on 14 May 2024).
39. Urcuioli, P.J. Stimulus Control and Stimulus Class Formation. In *APA Handbooks in Psychology®*; American Psychological Association: Washington, DC, USA, 2013; pp. 361–386. [CrossRef]
40. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **2023**, *55*, 194. [CrossRef]
41. Beeching, E.; Fourrier, C.; Habib, N.; Han, S.; Lambert, N.; Rajani, N.; Sanseviero, O.; Tunstall, L.; Wolf, T. Open LLM Leaderboard. 2023. Available online: https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard (accessed on 25 June 2024).
42. Gao, L.; Tow, J.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; McDonell, K.; Muennighoff, N.; et al. A Framework for Few-Shot Language Model Evaluation. 2021. Available online: https://zenodo.org/records/12608602 (accessed on 24 May 2024).
43. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457.
44. Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; Choi, Y. HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv* **2019**, arXiv:1905.07830.
45. Sakaguchi, K.; Bras, R.L.; Bhagavatula, C.; Choi, Y. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. *arXiv* **2019**, arXiv:1907.10641.
46. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. *arXiv* **2021**, arXiv:2009.03300.
47. Lin, S.; Hilton, J.; Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv* **2022**, arXiv:2109.07958.
48. Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. Training Verifiers to Solve Math Word Problems. *arXiv* **2021**, arXiv:2110.14168.
49. Chollet, F. On the Measure of Intelligence. *arXiv* **2019**, arXiv:1911.01547.
50. Santoro, A.; Hill, F.; Barrett, D.G.T.; Morcos, A.S.; Lillicrap, T.P. Measuring abstract reasoning in neural networks. *arXiv* **2018**, arXiv:1807.04225.
51. Dixon, M.R.; Belisle, J.; Stanley, C.R. Derived Relational Responding and Intelligence: Assessing the Relationship Between the PEAK-E Pre-assessment and IQ with Individuals with Autism and Related Disabilities. *Psychol. Rec.* **2018**, *68*, 419–430. [CrossRef]
52. Casper, S.; Davies, X.; Shi, C.; Gilbert, T.K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv* **2023**, arXiv:2307.15217.