Universiteit Leiden

Master Computer Science

Extending Cross-Modal Association Research: Investigating the Bouba-Kiki Effect in LLaMA and Molmo

Name: Robin R.P.M. Kras
Student ID: s4239008

Date: 2 July 2025

Specialisation: Data Science

1st supervisor: Tessa Verhoef
2nd supervisor: Rob Saunders

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Vision-language models (VLMs) represent a significant leap in the integration of multimodal learning in artificial intelligence, as they are designed to process and relate both visual and textual information. In this work, we investigate whether state-of-the-art multimodal AI models (Molmo and LLaMA3.2) exhibit human-like cross-modal associations. Specifically, this research addresses the research question of whether these models exhibit tendencies that align with a Bouba-Kiki effect, which is a well-known bias in human intuition and perception. This thesis tests both models robustly by using both probability-based matching tasks, probing model preferences, and employing a methods from the interpretability literature to analyse visual attention patterns. The results show that there is inconsistent, weak evidence in support of the Bouba-Kiki effect within these models, thereby enabling the conclusion that these models do not exhibit human-like cross-modal associations. All in all, this research aims to contribute to a deeper understanding of AI cognition and multimodal training, with implications for both model interpretability and human-AI interaction.

Contents

1 Introduction **5**
- 1.1 Motivation . . . 6
- 1.2 Contributions . . . 6
- 1.3 New Developments . . . 7

2 Background **8**
- 2.1 The Bouba-Kiki Effect in Human Cognition . . . 8
- 2.2 Cross-Modal Associations in AI & VLMs . . . 9
- 2.3 VLMs and Their Capabilities . . . 10
- 2.4 Attention Mechanisms & Model Explainability . . . 11
- 2.5 LLaMA3.2: A General-Purpose VLM . . . 11
- 2.6 Molmo: A Specialized VLM . . . 12

3 Methodology **13**
- 3.1 Method Overview . . . 13
- 3.2 Data and Models . . . 13
- 3.3 Prompt Engineering . . . 13
- 3.4 Experimental Setup . . . 15
- 3.5 Cross-modal Probability Analysis . . . 15
- 3.6 Image-to-Text Matching . . . 17
- 3.7 Attention Pattern Analysis . . . 18

4 Results **21**
- 4.1 Cross-Modal Probability Analysis . . . 21
  - 4.1.1 Congruency Effects . . . 21
  - 4.1.2 Effect Strength . . . 22
  - 4.1.3 Curved and Jagged Comparison . . . 24
  - 4.1.4 Congruent and Incongruent Probabilities . . . 25
  - 4.1.5 Significance Analysis . . . 26
- 4.2 Image-to-Text Matching . . . 27
  - 4.2.1 Overall Classification Scores . . . 27
  - 4.2.2 Score Difference . . . 28
  - 4.2.3 Sonorant-Rounded vs. Plosive non-Rounded . . . 29
  - 4.2.4 Significance Analysis . . . 31
- 4.3 Attention Pattern Analysis . . . 32
  - 4.3.1 Congruent vs Incongruent Shapes . . . 32
  - 4.3.2 Congruency Evaluation . . . 35
  - 4.3.3 Statistical Significance Analysis . . . 35

5 Summary of Results **36**

6 Discussion **38**
- 6.1 Interpretation of Congruency Patterns . . . 38
- 6.2 Comparing LLaMA and Molmo . . . 38
- 6.3 Implications for Cognitive Modeling . . . 39
- 6.4 Challenges in Measuring Cross-Modal Semantics . . . 40

7 Limitations **41**

8 Future Work **42**

9 Appendices **48**
9.1 Appendix A 48
9.2 Appendix B 50
9.3 Appendix C 59

1 Introduction

Large Language Models (LLMs) and Vision-and-Language Models (VLMs) have seen rapid advancements in recent years, leading to widespread adoption across industries, research fields, and everyday software applications like ChatGPT. Beyond their intended capabilities, these models occasionally exhibit unexpected and often unpredictable properties, referred to as emergent abilities [Wei et al., 2022]. These abilities, which arise as a byproduct of large-scale training, include advanced reasoning, zero-shot learning, and potential cognitive-like behaviors.

While some of these emergent abilities appear to be learnable from exposure to large-scale, human-generated text and image data, other cognitive traits depend more heavily on real-world, embodied experience. One such class of traits involves cross-modal associations, which are the tendency to link sensory modalities such as sound or vision in a structured way. These associations, which are deeply ingrained in human perception and cognition, raise the question of whether purely data-driven models can internalize such mappings without ever experiencing the world in a sensory, physical manner. This makes them a highly compelling test case for investigating the cognitive alignment of VLMs.

Cross-modal associations are particularly relevant to the study of human perception and AI alignment. A well-known example of this is the Bouba-Kiki effect, where humans tend to associate certain phonetic sounds with specific visual shapes, such as the word "bouba" with rounded shapes and "kiki" with jagged ones [Ramachandran and Hubbard, 2001, Köhler, 1929, Köhler, 1947]. A simple example of this effect and its corresponding sound-shape alignment to each word are shown in figure 1. These kinds of universally shared associations play an important role in the way how humans process and learn languages and likely partially dictate the way how communication systems are shaped [Fort et al., 2018].

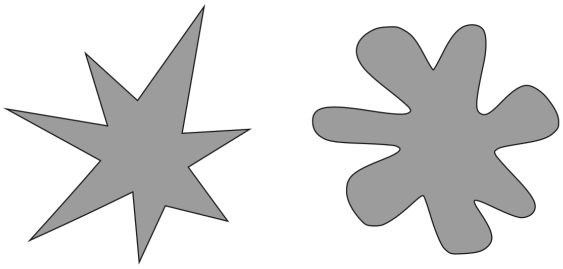

Figure 1: The Bouba-Kiki effect. Which image corresponds to Bouba and which image corresponds to Kiki? Images from Köhler (1929, 1947)

Previous research has resulted in a diverse set of outcomes, both positive (e.g., [Marklová et al., 2025]) and negative (e.g., [Verhoef et al., 2024] [Loakman et al., 2024] [Alper and Averbuch-Elor, 2023]) in supporting the idea that VLMs are capable of understanding human-like associations between cross-modalities.

This thesis extends prior work by investigating whether newer multimodal AI models - Molmo and LLaMA3.2, both released in Fall 2024 and representing state-of-the-art multimodal models - exhibit human-like cross-modal associations. Specifically, this research addresses the research question of whether LLaMA3.2 and Molmo exhibit Bouba-Kiki tendencies in a manner similar to human intuition and perception. Besides this, this research aims to validate how these models perform in probability-based matching tasks and how their focus regions are visualized and whether they fall in line with Bouba-Kiki associative expectations.

To answer these questions, this study will conduct 3 controlled experiments:

- Cross-modal probability analysis to determine how strongly models associate words with shapes.
- Image-to-text matching to evaluate model predictions and confidence levels for Bouba- and Kiki-pseudowords on both soft and sharp images.
- Attention pattern analysis using `Segment Anything Model V2 (SAM2)` to visualize focus regions in image-text tasks.

By systematically evaluating these models, this research aims to contribute to a deeper understanding of AI cognition and multimodal learning, with implications for both model interpretability and human-AI interaction.

1.1 Motivation

Given the importance of human cognitive biases and preferences like the Bouba-Kiki effect in shaping the way we communnicate, it is crucial to increase the understanding of cross-modal associations in AI to advance human-machine interaction and model interpretability. Hence, the Bouba-Kiki effect, a well-established phenomenon in human perception [Ramachandran and Hubbard, 2001, Köhler, 1929], has been explored in VLMs such as CLIP [Verhoef et al., 2024]. Although prior studies did not show strong evidence for the Bouba-Kiki effect in VLMs, CLIP and GPT-4o showed partial alignment with human preferences, therefore suggesting that further research on the topic using cutting-edge VLMs could prove something stronger. However, BLIP2 and ViLT did not show clear evidence of this effect, indicating the results to be inconsistent and inconclusive.

This research serves to fill this gap by systematically analyzing whether new, innovative models exhibit emergent cross-modal abilities, particularly in word-shape associations. In addition, by employing attention-based analytics, this research hopes to provide further insights into how these models process visual and linguistic information.

1.2 Contributions

This research will extend previous work by creating new and extending on previously conducted experiments.

- Investigating whether Molmo and LLaMA3.2 exhibit human-like cross-modal associations, specifically the Bouba-Kiki effect.
- Analyzing their attention patterns to determine whether the models exhibit visual focus on shape-relevant regions, similar to humans.
- Providing a comprehensive evaluation framework that can be applied to future multimodal research.

1.3 New Developments

Unlike previous research on VLMs ([Verhoef et al., 2024]), where CLIP, GPT-4o, BLIP2, and ViLT were the main topics in combination with cross-modal association testing, the main focus of this research will be to extend research to cutting-edge VLMs: Molmo and LLaMA3.2.

According to the developers, Molmo is a vision-language model that is part of a new generation of VLMs where openness remains central to the development and maintenance of the methodology. To address this, a new dataset called PixMo has been introduced [Deitke et al., 2024]. It includes highly detailed captions for pre-training, a free-form image Q&A dataset for fine-tuning, and an innovative 2D pointing dataset, all collected without external VLMs. As openness remains central to this methodology, research collaboration becomes more attractive, allowing a broader community to benefit from their knowledge and findings and advancing multimodal AI research.

In summer 2024, the Meta LLaMA team released its new herd of foundational LLaMA models, the LLaMA 3 herd. By publicly releasing these models, AI and LLM research has been greatly sped up, as these models include both pretrained and post-trained variants, thereby only necessitating downstream specification if used in public environments or settings. By keeping three factors central during the development process, including data, scale, and complexity, the Meta LLaMA team has managed to support the foundation of AI and LLM research through increasing the availability of knowledge on 'understanding key factors in foundational model development and contributing to a more informed debate about the future of foundational models for public use' [Grattafiori et al., 2024]. In this research, LLaMA3.2 will be explored and evaluated for its potential emergent capability of cross-modal association. The reason for opting for LLaMA3.2 over other variants is because LLaMA3.2 was the first model in the LLaMA herd that enabled the processing of text and images [Meta, 2024], which in turn enables the training and evaluation of cross-modal associations.

2 Background

Cross-modal associations can be understood as the way humans integrate information from multiple sensory modalities. Although some associations are commonly understood to be taught or learned, many others appear to be universal and biologically rooted within the nature of our species. In 1974, Lawrence E. Marks discovered that there was something that led to subjects associating auditory and visual brightness, albeit he could not understand what it was that led these subjects to do so [Marks, 1974]. Although this research is quite dated, this is still relevant for this newer age of research in cross-modal associations. AI models, specifically VLMs, aim to bridge the previously unseen gap between vision and language. Hence, researching cross-modal associations in VLMs will potentially increase the knowledge supply to advance the field of AI and VLMs in regards to whether these models are capable of mimicking human-like cognition or proving that they only rely on statistical shortcuts.

As coined by Köhler in 1929, the Bouba-Kiki effect (originally known as 'maluma' and 'takete' instead) is a suitable example to illustrate the connection subjects drew between spoken words and the shapes they would assign to them [Köhler, 1929]. This connection is included in figure 1. To support this connection, Ramachadran & Hubbard confirmed in 2001 that most people associate "bouba" with rounded shapes and "kiki" with jagged shapes [Ramachandran and Hubbard, 2001]. This effect has been observed across languages and cultures; Maurer et al. (2006) showed that both English-speaking children and Swahili-speaking children made similar associations despite differences in language and environment [Maurer et al., 2006].

This leads to the question of "if VLMs like CLIP, Molmo, and LLaMA3.2 are trained to align text and images, do they acquire human-like cross-modal associations?". There are several potential conclusions to this question. First, if a model exhibits the Bouba-Kiki effect, then it suggests that AI can learn abstract multimodal relationships. Second, if it does not exhibit the Bouba-Kiki effect, then it implies that AI relies on mechanisms different from humans to process cross-modal data. These two outcomes also have potential for broader implications, such as:

- Can AI models flexibly deal with the dynamic nature of human language in which new meanings continually emerge that are often rooted in abstract patterns, embodied metaphors or cross-modal preferences?
- Does cross-modal learning in AI improve or advance human-computer interaction and mutual understanding?

2.1 The Bouba-Kiki Effect in Human Cognition

The Bouba-Kiki effect highlights a robust cross-modal association between auditory and visual perception. This connection is often explained by phonetic properties: rounded sounds like /b/, /m/, and /o/ involve smooth articulation, whereas sharp sounds like /k/, /t/, and /i/ have abrupt, stop-like articulation [Nielsen and Rendall, 2012]. Researchers suggest that this psychological link is universal and prelinguistic, as it even appears in infants and non-verbal individuals [Maurer et al., 2006]. Furthermore, sound symbolism may support early language acquisition by strengthening connections between auditory and visual stimuli [Imai et al., 2008, Ćwiek et al., 2022].

The Bouba-Kiki effect, introduced informally by Kohler in 1929 [Köhler, 1929], was the effect that suggested an auditory and visual connection as a robust cross-modal association. The

experiment from this study was simple. As previously described: participants would connect a spoken word with a picture of a shape. The results fell in line with the expectations; most people associated ”bouba” with rounded shapes and ”kiki” with jagged shapes. Today, this effect is still understood to be universal and automatic by human nature, while also being consistent across cultures, languages, and writing systems [Ramachandran and Hubbard, 2001, Ćwiek et al., 2022]. The idea that humans match rounded sounds (e.g., /b/, /m/, /o/) with curved shapes is due to smooth articulation, whereas sharp sounds (e.g., /k/, /t/, /i/) match jagged shapes due to abrupt, stop-like articulation. Various studies show that these phonetic properties influence perception in languages [Nielsen and Rendall, 2012, Maurer et al., 2006]. This psychological connection is believed to be prelinguistic and universal considering that infants and non-verbal individuals exhibit the same effect [Maurer et al., 2006]. Additionally, some researchers argue that sound symbolism plays an important role in early language development [Imai et al., 2008], ultimately suggesting that stimulating the connection between auditory and visual impulses improves learning capabilities.

From a neurological perspective, the superior temporal sulcus is involved in integrating visual and auditory stimuli. The research proposed by Ramachandran and Hubbard suggests that cross-activation between the auditory and visual cortices explains why people naturally link sounds with shapes [Ramachandran and Hubbard, 2001]. In addition, further evidence in favor of this observation has been gathered by fMRI studies showing that the fusiform gyrus, which is the part of the brain responsible for shape processing, and the auditory cortex are both activated in Bouba-Kiki matching tasks [Peiffer-Smadja and Cohen, 2019].

Research on early language development shows that infants as young as four months old exhibit a Bouba-Kiki-like preference [Ozturk et al., 2013], suggesting that this phenomenon is preverbal rather than learned. Considering that this effect is observed across diverse languages and cultures [Maurer et al., 2006], the ability to connect sound and shape has likely been one of the consequences of evolution. For example, the ability to connect sound and shape may have helped prehistoric humans process threats, as sharp sounds indicate danger and soft sounds indicate friendly interaction. This aligns with theories on sound symbolism in early language evolution [Perlman et al., 2015]. Alternatively, it may also be due to common associations in our experiences with the physical world [Fort and Schwartz, 2022].

All in all, the Bouba-Kiki effect has broad implications on language and communication. It relates to language evolution, marketing and branding, speech therapy, and autism research. Therefore, it is beneficial to consider the Bouba-Kiki effect in vision-language models.

2.2 Cross-Modal Associations in AI & VLMs

Previous research has explored whether VLMs exhibit similar associations, with models such as CLIP showing inconsistent alignment with human perception [Verhoef et al., 2024]. Similarly, a recent study by Loakman et al. (2024) conducted controlled shape symbolism experiments using various VLMs and LLMs, including GPT-4 and LLaVA, and found that while some models could approximate human-like shape-sound associations, especially with additional task context, overall agreement with human judgments remained low and highly variable across models and conditions [Loakman et al., 2024]. Even more recent, Marklová et al. (2025) took a different approach by testing whether GPT-4 could generate pseudowords that carry iconic relationships between sound and meaning. They found that not only could LLMs produce pseudowords whose meanings could be guessed above chance by human participants, but also that the models themselves (GPT-4 and Claude3.5 Sonnet) outperformed humans in matching these words to

their intended meanings. This suggests that, in some cases, LLMs are capable of learning and applying non-arbitrary, cognitively grounded sound-meaning associations, despite a lack of pretraining or fine-tuning for the task at hand [Marklová et al., 2025]. However, given the conflicting results in prior work, more research is necessary to make any definitive statements.

Other early attempts in recreating the Bouba-Kiki effect in artificial intelligence with multimodal reasoning resulted in diverse outcomes. Alper and Averbuch-Elor (2023) hypothesized that sound symbolism is reflected in VLMs such as CLIP and Stable Diffusion ([Alper and Averbuch-Elor, 2023]). By employing zero-shot learning and probing, they investigated whether the innate ability and knowledge to draw multimodal connections was present in these models. From their research, they concluded that there is strong evidence to support this hypothesis.

Although this research suggests that VLMs do in fact have the ability to draw this connection, other research suggests that the evidence is too limited to confirm this conclusion. For example, Verhoef et al. (2024) experimented using a similar hypothesis to investigate whether the Bouba-Kiki effect was present in models CLIP, BLIP2, ViLT, and GPT-4o [Verhoef et al., 2024]. Ultimately, Verhoef et al. concluded that the evidence was too limited to draw any concrete statements or conclusions in favor of the existence of the Bouba-Kiki effect in VLMs.

2.3 VLMs and Their Capabilities

Vision-language models represent a significant leap in the integration of multimodal learning in artificial intelligence, as they are designed to process and relate both visual and textual information. As a consequence of the ability to analyse and generate content across modalities, these models have emerged as a promising venture to advance artificial intelligence and its capabilities. VLMs are typically trained on large amounts of paired data that link images and corresponding textual descriptions, enabling the learning of mappings between visual features and semantic contents. VLMs often rely on transformer architectures that process both text and images together. The most notable feature of transformer architectures is their ability to learn relationships and context from sequential data [Islam et al., 2023]. There are various methods to pre-train models that use the transformer architecture. Key approaches include:

- Contrastive learning (e.g., CLIP): Method that maximizes similarity between matching image-text pairs while minimizing mismatches [Radford et al., 2021];
- Masked language modeling (MLM): Method that predicts missing words in captions [Devlin et al., 2019];
- Image-text fusion models (e.g., BLIP2): Method that combines vision and text through early or late fusion layers [Li et al., 2023].

In addition to recognizing and generating visual-text mappings, VLMs also exhibit zero-shot learning, allowing them to classify novel images without explicit training. Moreover, they enable image-text retrieval, where text queries can locate relevant images in a dataset. For example, CLIP is able to recognize a description of ”a golden retriever playing outside” when given the image of a dog, even without fine-tuning on that example. In addition to zero- and few-shot learning capabilities, VLMs are also capable of common-sense reasoning and visual question answering (VQA). More specifically, advanced VLMs such as Flamingo and LLaVA are capable of answering questions about images, thus suggesting that they possess the ability to comprehend spatial relations.

Despite their impressive capabilities, VLMs still face significant challenges. One such challenge is bias in training data; since VLMs learn from internet-scale data, they inherit human-like biases, such as stereotypes and cultural misconceptions, thereby amplifying societal biases, including stereotypes related to gender, race, and culture [Bolukbasi et al., 2016, Bender et al., 2021]. Another challenge is spatial reasoning discrepancies. Many VLMs struggle with complex relationships that humans deem trivial, such as figuring out which object is in front of the other, or vice versa [Yatskar et al., 2016]. Lastly, the computational costs of training VLMs require massive GPU clusters and thousands of hours, thereby making these models very expensive and consequently inaccessible to smaller research teams [Strubell et al., 2020].

2.4 Attention Mechanisms & Model Explainability

Attention mechanisms are crucial in transformers and VLMs. The attention mechanism is defined as a method that helps models focus on the most relevant parts of input data. More specifically, attention mechanisms can be divided into three different types:

- Self-Attention: Models dynamically weigh different parts of input sequences.
- Cross-Attention: In multimodal models, text attends to image features and vice versa.
- Early vs. Late Fusion in VLMs: Some models combine modalities early, while others process them separately before merging outputs.

The transformer model is the first model to have been made exclusively using self-attention. Currently, self-attention has been successfully used in a variety of tasks, including ”reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations”, as confirmed by Vaswani et al. [Vaswani et al., 2017].

Attention mechanisms enable models like Molmo and LLaMA3.2 to selectively focus on important image and text features. However, despite their strength, attention does not always provide interpretable or human-like explanations. This issue is further highlighted by the work of Jain and Wallace (2019), who researched the explainability of attention mechanisms during decision-making processes [Jain and Wallace, 2019]. Since VLMs often act as black boxes, their decision-making is hard to interpret. Therefore, exploring the inner workings of the attention mechanism is crucial to ensure that models align with human perception. To investigate this, a variety of tools are usable, such as SAM2 and Grad-CAM. Both are techniques that helps visualize where VLMs “look”, although their decision-making process remains partially opaque [Selvaraju et al., 2017].

2.5 LLaMA3.2: A General-Purpose VLM

LLaMA3.2 (2024) bridges the gap between large-scale language models (LLMs) and VLMs by integrating visual perception with text-based reasoning. Unlike Molmo, which is optimized for image-text retrieval, LLaMA3.2 is designed for multimodal conversational AI and vision-language inference tasks, making it suitable for AI assistants [Meta, 2024].

Developed by Meta, LLaMA3.2 is an upgraded multimodal LLM, extending the capabilities of LLaMA3.1 by adding visual input processing, thereby positioning itself alongside similar models such as GPT-4o and Gemini1.5. Previous studies have explored the multimodal capabilities of GPT-4o [Verhoef et al., 2024], and similar investigations into the performance of LLaMA3.2

on cross-modal tasks will provide valuable insights into its potential to advance cross-modal associative research.

2.6 Molmo: A Specialized VLM

Molmo (2024) represents an important shift towards open-access VLMs. By allowing researchers to explore cross-modal learning without the restriction of proprietary models, Molmo has the potential to greatly advance VLM research. Developed by Ai2, Molmo is an open-source VLM designed as an alternative to proprietary systems [Deitke et al., 2024].

Trained on the massive multimodal dataset called PixMo, it excels in image-text retrieval, captioning, and reasoning tasks. PixMo has been a driving force in the development of Molmo as it contains data collected from human annotators using speech-based descriptions. Since the Bouba-Kiki effect is often explained on the basis of speech sounds and articulatory constraints, it makes Molmo an excellent candidate for the investigation of this effect. Molmo has been developed using contrastive learning, similar to CLIP. Moreover, it supports fine-tuning for domain-specific applications. Despite its similarities to CLIP, Molmo's architecture and weights are fully available online, making it an ideal candidate for potential further exploration of its inner workings.

3 Methodology

3.1 Method Overview

This section describes the methodology used to investigate cross-modal associations in vision-language models LLaMA3.2 and Molmo. The goal is to determine whether these models exhibit the Bouba-Kiki effect by evaluating their performance across 3 distinct experimental setups:

1. Cross-Modal Probability Analysis – Tests how strongly models associate certain words with certain images.

2. Image-to-Text Matching – Evaluates the confidence level that is assigned to Bouba- and Kiki-pseudowords based on certain images.

3. Attention Pattern Analysis – Uses `Segment Anything Model V2 (SAM2)` to visualize how models focus on image regions.

Each experiment is designed to probe a different aspect of how VLMs process and associate images with text, providing insight into their alignment with human cognitive biases.

3.2 Data and Models

The dataset used in this study consists of images representing both curved and jagged shapes, sourced from prior research on the Bouba-Kiki effect [Verhoef et al., 2024] and augmented with synthetic shapes generated by using simple geometric transformations. The accompanying word set includes sonorant-rounded and plosive non-rounded Bouba-Kiki pseudowords, sourced by previous research from Nielsen and Rendall, [Nielsen and Rendall, 2012] and previously experimented with by Verhoef et al. [Verhoef et al., 2024]. Both of these sets can be found in appendix C 9.3 and A **??**, respectively.

The evaluated models are Molmo and LLaMA3.2, two state-of-the-art VLMs. These models are chosen because of their capability to process both textual and visual inputs with a shared embedding space, as well as their ability to perform visual question answering. Before experimentation, all input data undergo preprocessing, including resizing images to a consistent dimension, normalizing pixel values, and tokenizing inputs using the respective model's tokenizer. Furthermore, in order to enhance the reproducibility of this research, we recommend using variants `meta-llama/LLaMA-3.2-11B-Vision-Instruct` [1] and `allenai/Molmo-7B-D-0924` [2].

3.3 Prompt Engineering

Prompts play a crucial role in shaping the outputs of vision-language models. Since these models rely on textual input to generate or interpret visual content, the way a question or instruction is phrased can significantly impact the response. A well-structured prompt can guide a model toward more accurate or meaningful outputs, whereas vague or poorly designed prompts may lead to inconsistent or misleading results. More specifically, for Molmo and LLaMA3.2, it became increasingly meaningful to generate prompts that both models accepted.

[1] `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct`

[2] `https://huggingface.co/allenai/Molmo-7B-D-0924`

Therefore, understanding the sensitivity of VLMs to prompt phrasing is essential, particularly when analyzing cross-modal associations such as those observed in the Bouba-Kiki effect.

Although the quality of the prompts themselves does not necessarily increase the learning ability of the models [Webson and Pavlick, 2022], minor differences in prompt wording can lead to substantial differences in model behavior. For example, a VLM tasked with identifying the relationship between a rounded shape and a given word may respond differently depending on whether a prompt is neutral or leading. As an example:

1. Neutral prompt: "Describe the relationship between this shape and the word 'Bouba'."

2. Leading prompt: "Does this rounded shape resemble the word 'Bouba' more than 'Kiki'?"

The first prompt is open-ended, thus allowing for a wider range of output, whereas the second prompt forcefully limits the model to reply in binary terms. Such variations can introduce bias into model evaluations, making it necessary to generate both normalized and specific prompts for the same purpose, thus implying that for this research, the prompts will follow two standard formulas specific to LLaMA3.2 and Molmo.

After experimenting it became apparent that LLaMA3.2 requires a very specific leading prompt to be evaluated accordingly, whereas Molmo had the capability to handle a variation of both neutral and leading prompts for these experiments. Hence, this study exclusively used leading prompts for both models to enhance the reliability and robustness of the framework and its outputs. Table 1 highlights the importance of prompts during the first experiment.

Prompt Type	Model	Example Prompt	Effectiveness
Leading	LLaMA3.2	Does this image represent the abstract word '{text_prompt}'?	Inconsistent
Leading	LLaMA3.2	Does this rounded shape resemble the word 'Bouba' more than 'Kiki'?	Inconsistent
Reinforced	Molmo	You are a classifier that responds with yes or no. Does this word '{text_prompt}' fit the image?	Inconsistent
Leading	LLaMA3.2	If you had to describe the following image with the abstract word '{text_prompt}', would you say it fits? Reply with only yes or no.	Semi-Consistent, Not Ideal
Neutral	LLaMA3.2 and Molmo	**On a scale from 0 to 100, how well does the abstract word '{text_prompt}' describe the following image? 0 means the word and image have no connection and 100 means the word perfectly fits the image. Answer only with a single number between 0 and 100. Do not include any other words.**	**Highly-Consistent, Contextually Ideal**

Table 1: Effectiveness of Different Prompt Types in Eliciting Expected Bouba-Kiki Responses During Experiment 1. '{text_prompt}' corresponds to the Python text variable from the list of Bouba-Kiki pseudowords.

The results in the table highlight that LLaMA3.2 requires delicate prompting, significantly tweaking the existing example prompt that was provided by Meta, whereas Molmo functions in a more simplistic manner, where only slight tweaking of the example prompt provided in the documentation on Hugging Face [3] was already sufficient.

[3] https://huggingface.co/allenai/Molmo-7B-D-0924

3.4 Experimental Setup

For the experiments, a variety of different settings have been experimented with and evaluated. The model outputs were generated using a balanced sampling configuration, included in table 2, which has been optimized to control variability. This specific configuration helps ensure that the model responses remain grounded and interpretable, while still allowing for some creative freedom. This is crucial for capturing subtle congruency effects in cross-modal associations. After comparing the preliminary model outputs, it became apparent that the balanced setting would be contextually ideal for this research, as it was appropriate for both models and could therefore be used as a middle ground. The temperature value of 0.7 introduced slight randomness within model generation, and although it was able to keep the focus of the task relatively well, a few invalid, non-numeric outputs occurred in the first experiment. By enabling the sampling mode, the models were allowed to express variability, which was kept on for all experiments to increase the reliability of the outputs. Nucleus sampling (top_p) was set to 0.9 to restrict token choices to the smallest groups of top-scoring tokens whose total probability was at least 90%, which ensured that the outputs were contextually relevant and applicable to both models. Lastly the top_k was set to 40 to narrow down the set of possible tokens to the 40 most-likely options, which prevents uncommon or off-topic words from being sampled during the experiments.

For reproducibility, generalizability, and increased reliability, the balanced setting has been repeated for all experiments. Moreover, all results for experiment 1 and 2 have been averaged over 10 iterations to increase reliability.

The data used for experiments 1, 2, and 3 consisted of 34 images containing an equal number of both curved and jagged shapes, whereas the dataset of the 3rd experiment consisted of 18 concatenated images, of which the curved and jagged shape are put alongside each other.

Mode	Temperature	do_sample	Top-p	Top-k	Description
Deterministic	0.6	True	0.85	40	Lower diversity; more focused and stable responses
Balanced	0.7	True	0.9	40	Moderate diversity; balance between focus and variety
Creative	0.8	True	0.95	50	Higher diversity; encourages exploratory responses

Table 2: Sampling configurations used for probing LLaMA3.2 & Molmo responses.

3.5 Cross-modal Probability Analysis

The first experiment evaluates whether vision-language models (VLMs), specifically LLaMA3.2 and Molmo, can detect congruency between abstract pseudowords and geometric shapes. The core functionality is implemented in a class-based Python framework, titled `CrossModalAnalyzer`, which automates data preparation, multimodal prompting, model inference, and statistical evaluation. More specifically, given an image of either a curved or jagged shape, the model is iteratively presented with a text prompt drawn from a set of four Bouba-Kiki pseudowords, where each prompt follows a set template:

Prompt:

```
"On a scale from 0 to 100, how well does the abstract word '{
   TEXT_PROMPT}' describe the following image? 0 means the word
    and image have no connection and 100 means the word
   perfectly fits the image. Answer only with a single number
   between 0 and 100. Do not include any other words."
```

The model is expected to return only a numeric score, which is parsed from its output. This score represents the certainty of the model's association between the word and the given image.

Class Design and Workflow

The `CrossModalAnalyzer` class is initialized using a folder containing images labelled as either "curved" or "jagged", respectively corresponding to "bouba/maluma" and "kiki/takete".

Once initialized, the class supports the following main steps:

1. **Model Loading**: The model is loaded through HuggingFace, using the processor and model checkpoints available through the platform.
2. **Image Scanning**: The image folder is scanned to extract valid images and their inferred types.
3. **Dataset Preparation**: All possible combinations of input images and 4 randomly sampled pseudowords (2 bouba-like, 2 kiki-like) are constructed. Each pair is annotated with metadata: image type, pseudoword type, and congruency.
4. **Multimodal Prompting**: For each image-word pair (36 images, 34*4 = 136 total iterations), the model receives the prompt specified above accompanied by the image. Next, the model returns a numeric output ranging from 0 to 100, which represents the perceived semantic fit between the image and pseudoword. The output is then parsed and validated, and invalid or non-numeric answers are logged as a NaN value for error-handling purposes.
5. **Metric Computation**: After all trials are complete, the class computes the following metrics:
    - Average congruent vs. incongruent score.
    - Bouba-Kiki effect strength (congruent minus incongruent average).
    - General bias (deviation from a 50-point neutral baseline).
    - Granular averages by image and pseudoword type (e.g., curved image + kiki word and vice versa).
    - Per-image score variance and standard deviation.
6. **Output Export**: The results, summary metrics, and per-image statistics are exported to .CSV files for further analysis. Steps 4-6 are repeated 10 times, using the same 4 pseudowords, to increase the reliability of the results.

The analyzer class for Molmo is set to operate in batches of four words and clears the GPU cache between batches to manage memory efficiently due to computational restraints, something that Molmo evidently required.

The results are aggregated over multiple runs to increase statistical reliability by reducing randomness or variability. This goes hand-in-hand with the balanced parameter options discussed in table 2.

In earlier versions of the experiment, the model was asked to respond with 'yes' or 'no' to assess the strength of the association. The final probability score was then calculated by applying a softmax on the logits for 'yes' and 'no', treating the result as a binary classification. However, after revision, it became apparent that this methodology would result in unreliable or lacking conclusions.

To create a realistic test environment and to limit the chance that the model was aware of the prior Bouba-Kiki experiments, every image is tested using two randomly sampled Bouba and Kiki pseudowords (4 in total). These words were: *'mohloo', 'lahmoo', 'kuhtuh'*, and *'taypay'*. Keeping the same four words has various benefits, including the reduction of linguistic variability and allowing for a more focused assessment of congruency effects. The lists of Bouba and Kiki pseudowords have been reused from prior research [Verhoef et al., 2024], and can be found in appendix A **??**.

Overall, this experiment provides a scalable and interpretable pipeline to assess whether LLaMA3.2 and Molmo exhibit cross-modal congruency in large-scale, cutting-edge VLMs. It further enables direct comparison between human-like associative biases and the semantic behavior of state-of-the-art VLMs. Moreover, it also enables investigation whether these patterns emerge despite the abstract nature of the input pseudowords and provides a method of evaluation on how a cutting-edge VLM interprets semantic fit between language and vision in a zero-shot setting.

3.6 Image-to-Text Matching

To evaluate cross-modal congruency between visual shapes and pseudowords, we implemented an image classification system using Meta's `LLaMA3.2` and AllenAI's `Molmo`. The classification process aimed to assess how likely an image of a shape (either curved or jagged) would be labelled using two pseudowords from two phonetically distinct categories: sonorant-rounded (S-R) and plosive non-rounded (P-NR).

Two distinct pseudoword label sets were randomly sampled from larger predefined pools, each containing 81 carefully constructed pseudowords per category. Ten unique pseudowords were randomly sampled on every iteration to ensure a manageable prompt size and balanced comparison for classification.

In each trial, the model is shown an image of a shape and asked to choose the most appropriate label. This is executed twice, once with a word from a randomly sampled subset of S-R and once with a set randomly sampled subset of P-NR pseudowords. The model output is recorded along with the confidence score based on the token-level probabilities returned by the model-specific decoder. Based on the returned pseudoword and accompanied confidence level, we analyzed whether these scores reflected congruent sound-shape mappings. Model performance is quantified using the accompanied label and confidence score.

Class Design and Workflow

The classification was performed using the following procedure specified within the `ImageTextMatcher` class. Once the authorisation of HuggingFace and GPU set-up is complete, it follows the following steps:

1. **Model Loading**: The model is loaded along with its associated processor.
2. **Initialisation**: For each input image, construct a chat-style prompt instructing the model to assign a label from the given lists.
3. **Classification**: Each image is classified twice, once with the S-R label set and once with the P-NR label set.
4. **Output Analysis**: The predicted label inside the output and the average confidence score derived from the softmax probability of each generated token are stored.

The instruction prompt for each image is inputted twice, once for each distinct label set. The prompt follows the structure of:

```
"You are given an image for which you need to assign a label.
   Use one of the following labels: '{LIST}'. Only respond with
    the label."
```

The classifier is set to operate in batches of four images and clears the GPU cache between batches to manage memory efficiently. Classification results are stored in a structured format that includes image path, predicted labels, confidence scores, and image type.

Lastly, a summary analysis computes the mean confidence scores of each label set across the curved and jagged images, which enables a quantitative comparison of the model its congruency preferences, which specifies whether the model is more confident in using S-R labels for curved images and P-NR labels for jagged ones, which ideally falls in line with the Bouba-Kiki effect.

After the execution of the program, it becomes possible to assess the scores for the distinct word sets, consequently enabling validation of whether the scores of S-R words are generally higher than the scores of P-NR words on their congruent image shape, and vice versa. If there exists a recurring trend within this comparison that falls in line with the expected result according to the Bouba-Kiki effect, then it implies a cross-modal connection in VLMs.

3.7 Attention Pattern Analysis

In this experiment we investigated the visual grounding and referential focus of the multimodal language model of `Molmo` in a forced-choice image comprehension task. This experiment aimed to determine whether Molmo's predicted spatial attention corresponded more closely to the curved or jagged side of a bipartite image, depending on the shape-sound congruency of the prompted pseudoword.

To this end, we reused the concatenated images from previous research [Shahrasbi, 2024], where the left and right halves featured distinct shapes: one soft and one jagged. We used the following pseudowords across all concatenated images: *mahnoo, lohmah, teepuh* and *kaytay* as prompts, chosen to reflect varying degrees of sound-symbolic congruency with the depicted shapes (e.g., mahnoo is more congruent with soft shapes, whereas kaytay is more congruent with jagged shapes.

Please note that this experiment has been performed exclusively on Molmo due to a lack of resources to perform this experiment on LLaMA, as LLaMA is a text-only language model, thus resulting in a lack of native support for image inputs or visual grounding.

The concatenated images follow the structure included in figures 2 and 3.

Figure 2: Concatenated image: curved shape vs. jagged shape.

Class Design and Workflow

The pipeline involved the following steps:

1. **Text and Image Input**: The text and image input are jointly processed using Molmo's associated processor, with prompts in the form of:

```
"Point to the '{PSEUDOWORD}' in the image."
```

2. **Inference**: The inference is run via the `generate_from_batch` function, and Molmo's response was parsed for spatial coordinate predictions using regular expressions to extract X and Y values embedded in the SVG-style output.

3. **Mappings**: These predicted coordinates were then mapped back to the pixel space based on the original image size.

To visualize and verify the semantic segmentation and attention localization:

- We used the `Segment Anything Model v2 (SAM2)` with a pre-trained `sam2_hiera_large` checkpoint to segment the region corresponding to the model predicted point.
- These segmented regions were overlaid on the input image, allowing us to assess whether the attention fell predominantly on the curved or jagged side of the concatenated image stimuli.

The input data for this experiment consists of 9 different images, of which each image has been mirrored once, thereby resulting in a set of 18 images. The focus point of the model is highlighted in blue, which can be seen in figure 3.

Figure 3: Molmo focus points using the prompt ”Point to Lohmah in the picture”. The star and the highlighted area in blue indicate the area that Molmo associates with the given prompt.

This experiment enabled a qualitative and semi-quantitative assessment of multimodal referential grounding by comparing segmented attention maps across different pseudowords and shape pairings.

4 Results

This section contains the results of experiments 1, 2, and 3: cross-modal probability analysis, image-to-text matching, and attention pattern analysis.

4.1 Cross-Modal Probability Analysis

This experiment investigates whether LLaMA3.2 and Molmo exhibit cross-modal congruency effects similar to human perceptual tendencies, such as preferring soft pseudowords for curved shapes and sharp pseudowords for jagged ones.

4.1.1 Congruency Effects

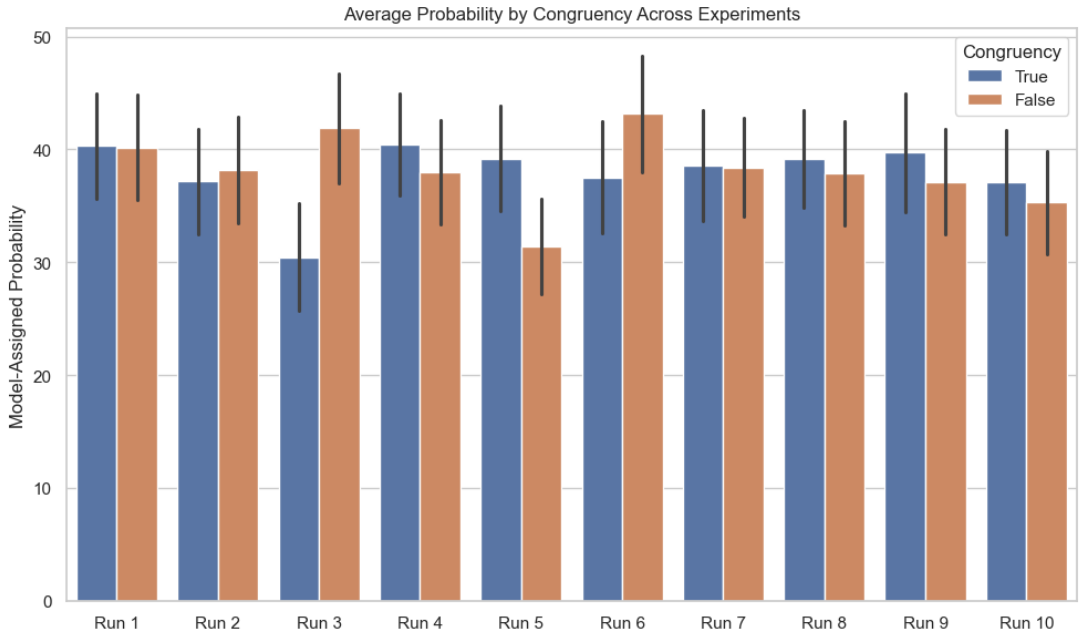

Figure 4: **LLaMA**: Assigned probabilities for all congruent and incongruent pairings across all iterations.

As shown in figures 4 and 5, the models rarely assigned higher probabilities to congruent pairings over incongruent pairings. In figure 4, LLaMA correctly assigned notably higher probabilities to congruent pairings in run 5, whereas run 3 highlights a strong contrast in which the incorrect congruent pairings received higher probabilities. These notable iterations support the argument that LLaMA does not consistently assign higher probabilities to congruent pairings, and thereby is not capable of finding the correct cross-modal connections. Similarly, in figure 5, the difference in probabilities between incongruent and congruent pairings was smaller, albeit displaying similar behavior to LLaMA in terms of inconsistently assigning higher probabilities to congruent pairings than to incongruent pairings. However, the discrepancies in performance, where there is no consistent trend of the congruent pairings receiving higher probabilities, suggests that the overall findings do not show reliable or stable proofs for the Bouba-Kiki effect in both LLaMA and Molmo.

Another trend, is that Molmo assigned significantly higher probabilities overall, ranging between $< 65, 73 >$, as opposed to LLaMA's $< 31, 44 >$ range. Since the average incongruent

probabilities remained lower than the congruent probabilities, it may suggest a statistically meaningful cross-modal correspondence, implying that Molmo is more aligned with humans in its decision-making than LLaMA with respect to the Bouba-Kiki domain.

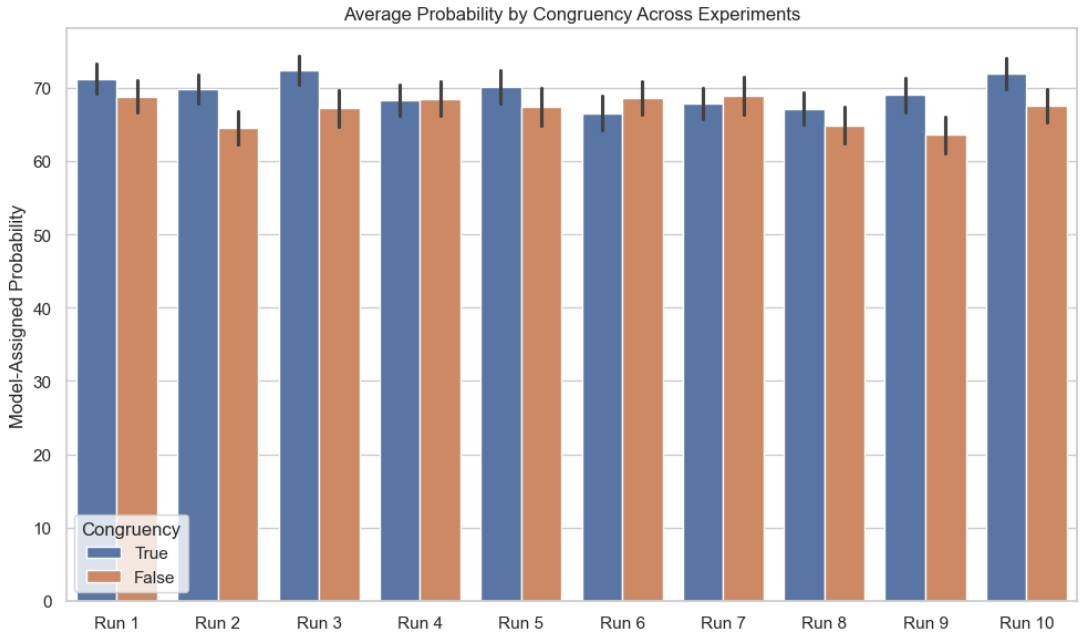

Figure 5: **Molmo**: assigned probabilities for all congruent and incongruent pairings across all iterations.

4.1.2 Effect Strength

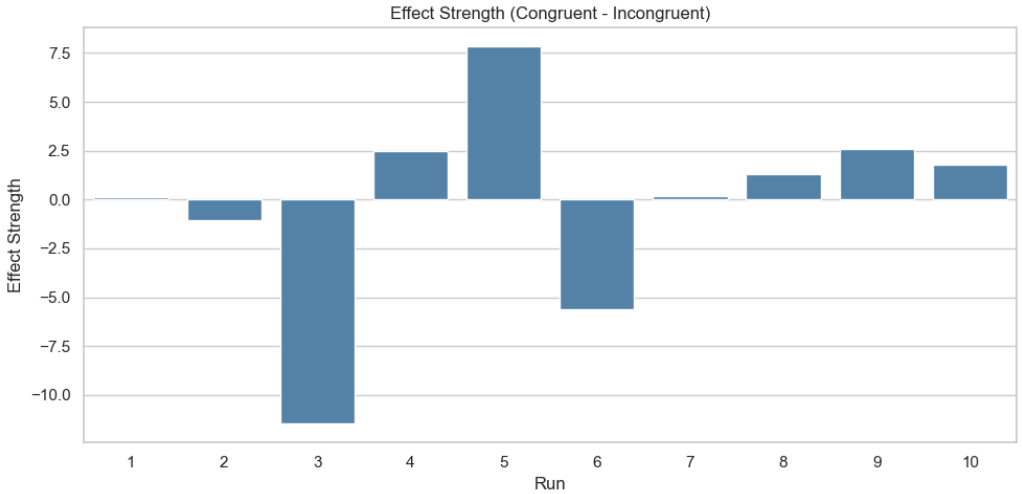

Figure 6: **LLaMA**: Effect strength between congruent and incongruent pairings. Effect strength denotes the difference between the assigned probabilities to congruent and incongruent pairings.

Figures 6 and 7 illustrate the computed effect strength, highlighting the difference between congruent and incongruent probabilities, and signified by the formula:

$$E = \Delta P_{congruent} - P_{incongruent}$$

Where E denotes the difference in probability between congruency and incongruency, with positive values observed in 6 out of 10 iterations for LLaMA and 7 out of 10 iterations for Molmo.

These values suggests that the models, specifically Molmo, exhibit slight sensitivity to shape-sound congruency. However, for LLaMA, the mean effect strength is negative, indicating no evidence for the Bouba-Kiki effect for this model. In contrast, Molmo has a occasional positive difference between congruency and incongruency, suggesting that it may be able to establish connections between congruent pairings, providing some support for the Bouba-Kiki hypothesis in this model, regardless of some inconsistencies per run.

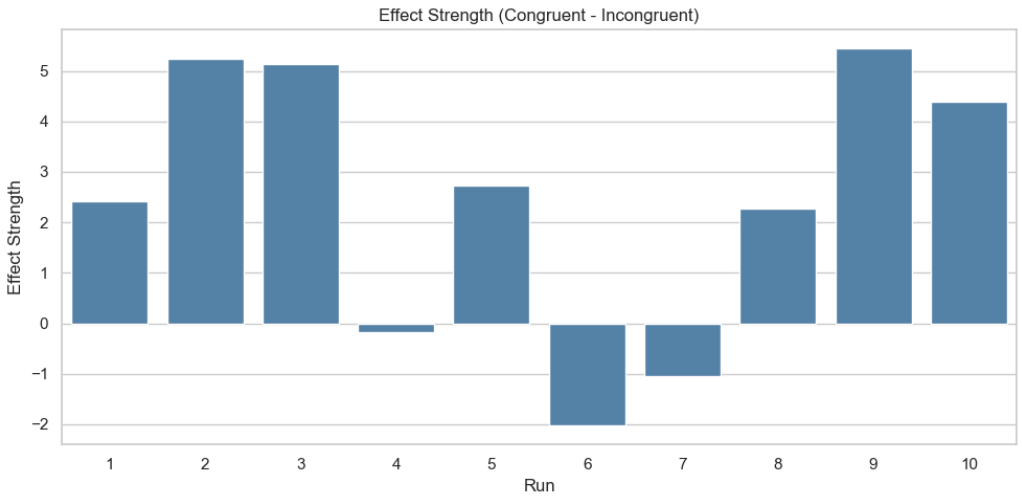

Figure 7: **Molmo**: Effect strength between congruent and incongruent pairings. Effect strength denotes the difference between the assigned probabilities to congruent and incongruent pairings.

4.1.3 Curved and Jagged Comparison

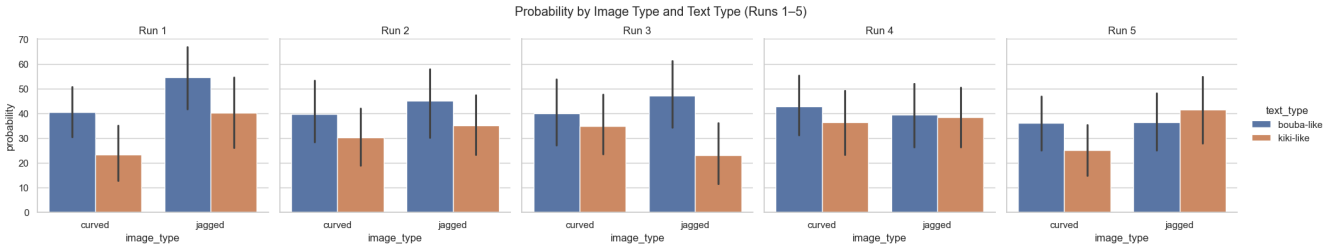

Figure 8: **LLaMA**: Assigned probabilities per image type (curved vs. jagged) (runs 1-6).

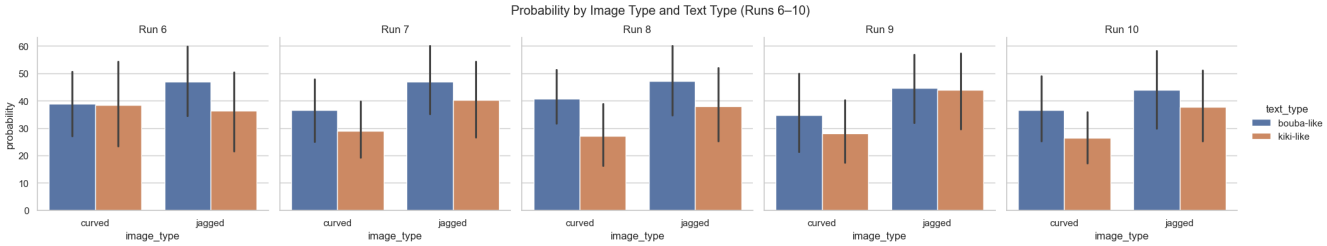

Figure 9: **LLaMA**: Assigned probabilities per image type (curved vs. jagged) (runs 6-10).

Figures 8, 9, 10, and 11 compare congruent probabilities across shape categories. Figures 8 and 9 show that, across all iterations, both curved and jagged images elicited higher probabilities for sonorant-rounded words than for plosive non-rounded words. This suggests that LLaMA tends to preferably associate both curved and jagged images with bouba-like pseudowords.

In contrast, figure 10 and 11 reveal that, across all iterations, both curved and jagged images elicited favor towards sonorant-rounded pseudowords, thus indicating that Molmo struggled to establish clear associations between sharp words and jagged shapes. Overall, these results indicate that neither LLaMA or Molmo demonstrated consistent congruency patterns, further supporting the lack of strong evidence for the Bouba-Kiki effect.

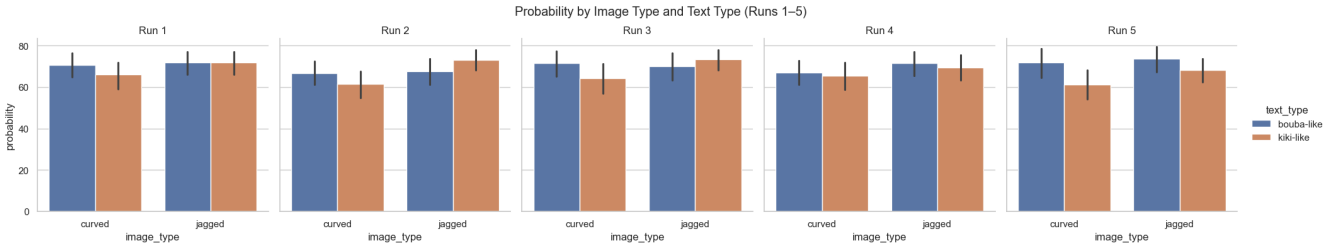

Figure 10: **Molmo**: Assigned probabilities per image type (curved vs. jagged) (runs 1-6).

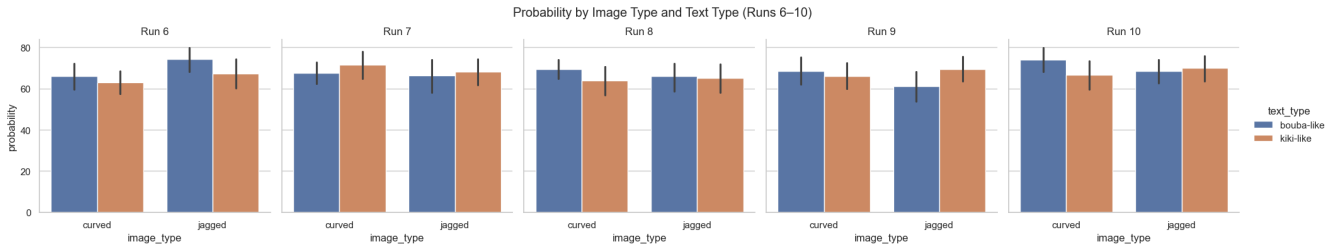

Figure 11: **Molmo**: Assigned probabilities per image type (curved vs. jagged) (runs 6-10).

4.1.4 Congruent and Incongruent Probabilities

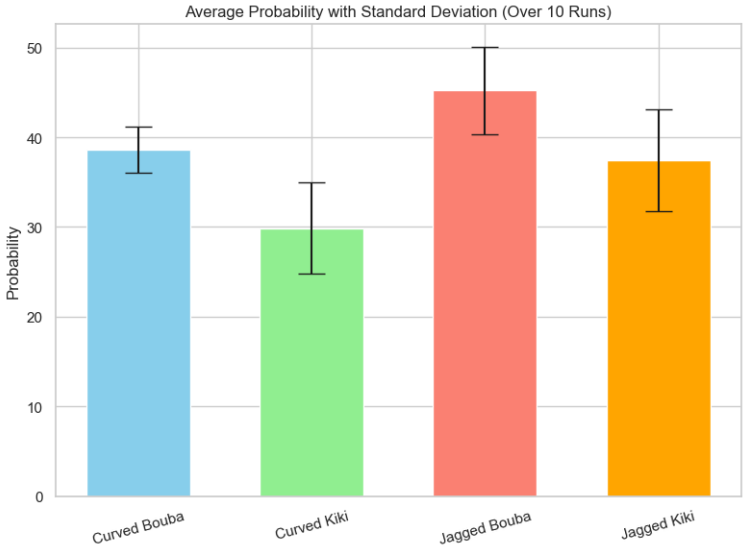

Figure 12: **LLaMA**: Assigned scores per image type (curved vs. jagged).

The results in figure 12 indicate that LLaMA exhibits an inconsistent pattern when assigning higher probabilities to congruent pairings compared to incongruent ones. Additionally, there is a clear tendency for sonorant-rounded pseudowords to receive higher scores than plosive non-rounded ones, regardless of the accompanying image. This suggests a potential bias towards softer-sounding words, independent of visual congruency.

In contrast, figure 13 shows that Molmo demonstrates a more balanced yet still inconsistent pattern in scoring congruency higher than incongruency, which could warrant further investigation.

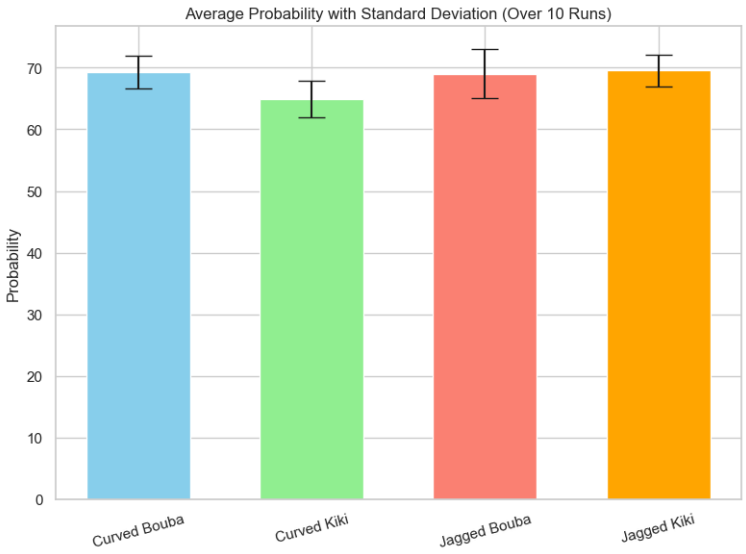

Figure 13: **Molmo**: Assigned scores per image type (curved vs. jagged).

4.1.5 Significance Analysis

Test	p-value	t-value	Congruent Mean	Incongruent Mean
Welch's t-test	0.946	-0.067	38.002	38.141

Table 3: Statistical significance test (Welch's t-test) results for LLaMA.

To assess statistical significance of these findings in LLaMA and Molmo, we have conducted a Welch's t-test. This choice was made based on the reason that the mean probabilities of congruency and incongruency do not follow a normal distribution. The results of these tests are included in tables 3 and 4.

From these results we can conclude that the the difference in probabilities between congruent (38.002) and incongruent (38.141) mappings in LLaMA is statistically insignificant.

On the other hand, the results of Molmo indicate a p-value of 0.017 and higher mean probability for congruent mappings (69.426) compared to incongruent mappings (66.989): a small, yet significant difference. These results support the hypothesis that Molmo exhibits sensitivity to cross-modal congruency, while LLaMA does not.

Test	p-value	t-value	Congruent Mean	Incongruent Mean
Welch's t-test	0.017	2.385	69.426	66.989

Table 4: Statistical significance test (Welch's t-test) results for Molmo.

4.2 Image-to-Text Matching

This experiment aimed to assess whether LLaMA and Molmo show higher confidence levels for Bouba- and Kiki-pseudowords on curved and jagged imagery.

4.2.1 Overall Classification Scores

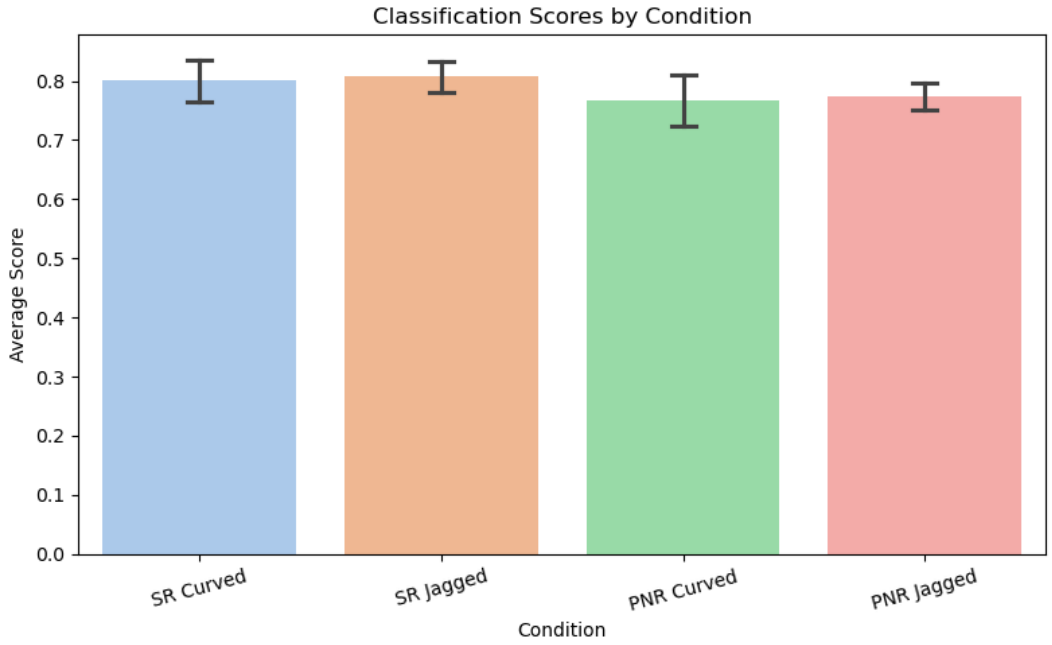

Figure 14: **LLaMA**: Average scores per word type across all 10 iterations.

Figures 14 and 15 highlight the obtained scores for both bouba-like and kiki-like pseudowords on curved and jagged images. The purpose of this figure is to indicate whether, on average, sonorant-rounded words score higher on curved images than plosive non-rounded, and vice versa.

The figures suggest that sonorant-rounded pseudowords obtain similar scores on both curved and jagged images. Moreover, plosive non-rounded pseudowords also score similar results across both image types. The only remarkable observation is that sonorant-rounded pseudowords obtain higher scores than plosive non-rounded pseudowords, no matter the image type.

The results of Molmo indicate identical behavior to LLaMA, where sonorant-rounded pseudowords obtain a marginally lower score higher on curved images than plosive non-rounded pseudowords. A similar, unexpected trend can be observed between sonorant-rounded and plosive non-rounded pseudowords on jagged images, thereby supporting an argument that these results do not deliver any support in favor of the Bouba-Kiki hypothesis.

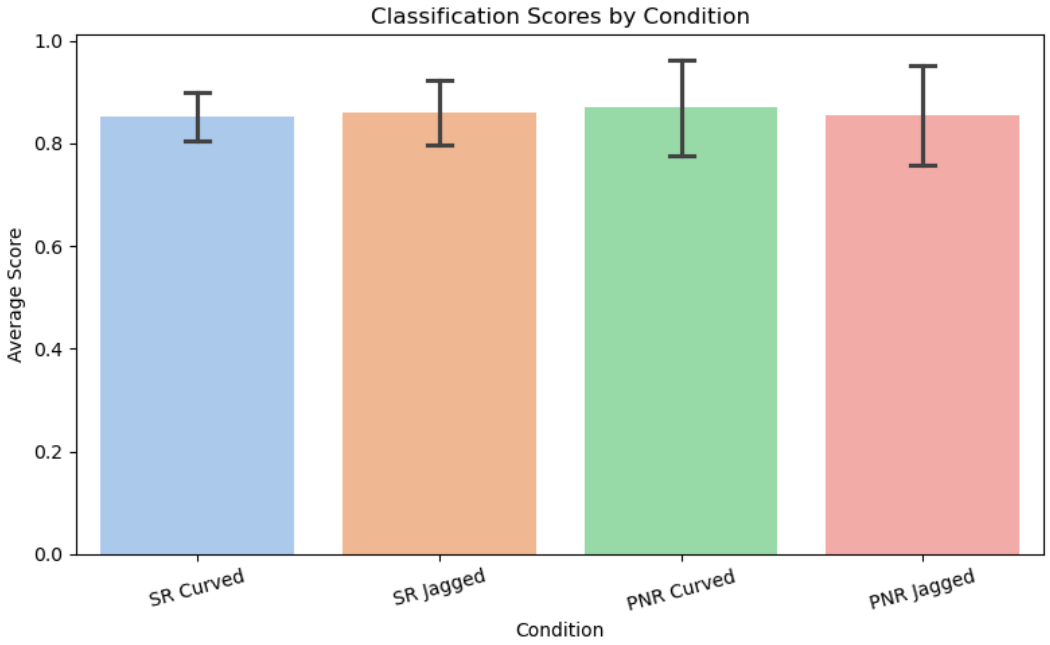

Figure 15: **Molmo**: Average scores per word type across all 10 iterations.

4.2.2 Score Difference

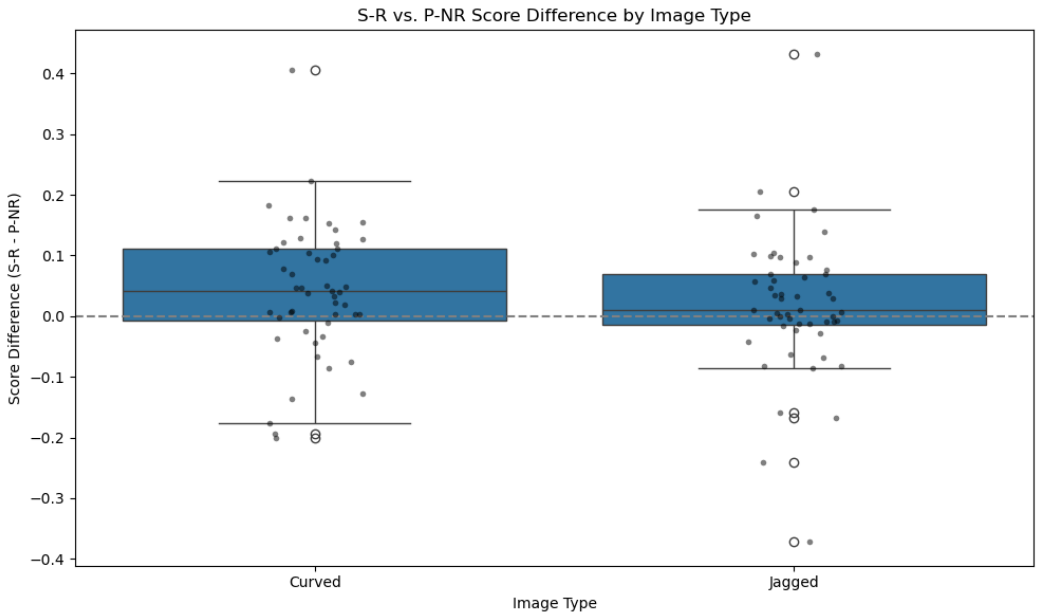

Figure 16: **LLaMA**: S-R and P-NR score difference across curved and jagged images.

Figures 16 and 17 display the score differences between sonorant-rounded pseudowords paired with curved images and plosive non-rounded pseudowords paired with jagged images. In both models, the score differences are insignificant, suggesting that sonorant-rounded pseudowords generally receive higher scores regardless of the accompanying image type. Ideally, we would expect that higher scores are assigned to sonorant-rounded pseudowords with curved images, and lower scores are assigned for plosive non-rounded pseudowords with jagged images. However, the lack of a substantial difference in the plots suggests limited evidence for this pattern.

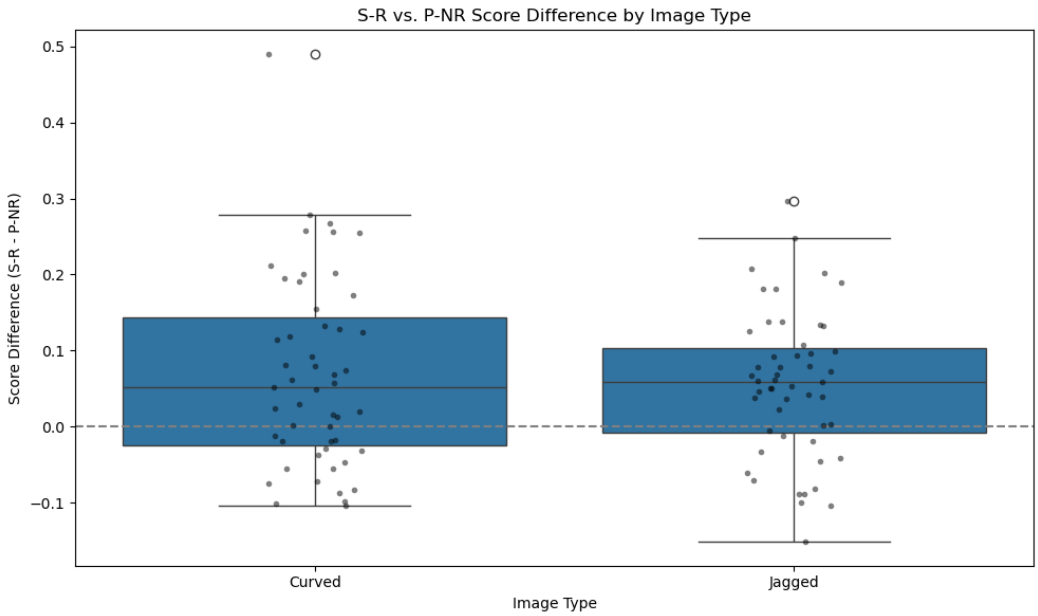

Figure 17: **Molmo**: S-R and P-NR score difference across curved and jagged images.

4.2.3 Sonorant-Rounded vs. Plosive non-Rounded

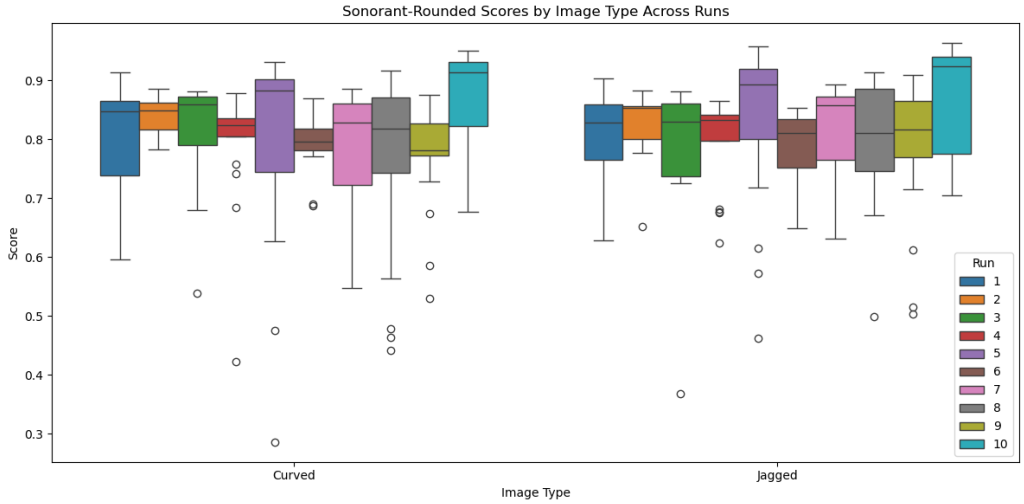

Figure 18: **LLaMA**: Average scores for Sonorant-Rounded words across all 10 iterations.

Figures 18 and 19 illustrate the differences in confidence score assigned to sonorant-rounded words when paired with curved versus jagged images, returned by both models.

Figure 18 shows that LLaMA assigns nearly identical confidence scores to S-R words for both curved and jagged images, with only minor variations. Given the minimal difference, it is difficult to conclude whether LLaMA has captured any meaningful association between sonorant-rounded pseudowords and curved images, thereby providing no supporting evidence in favor of the Bouba-Kiki effect.

In contrast, figure 19 indicates higher variability between confidence levels for sonorant-rounded pseudowords on curved and jagged images. However, Molmo inconsistently assigns higher confidence scores to sonorant-rounded pseudowords on curved images than on jagged

ones, thereby contradicting the expected bouba-like association. As a result, this model also fails to provide evidence for the Bouba-Kiki effect.

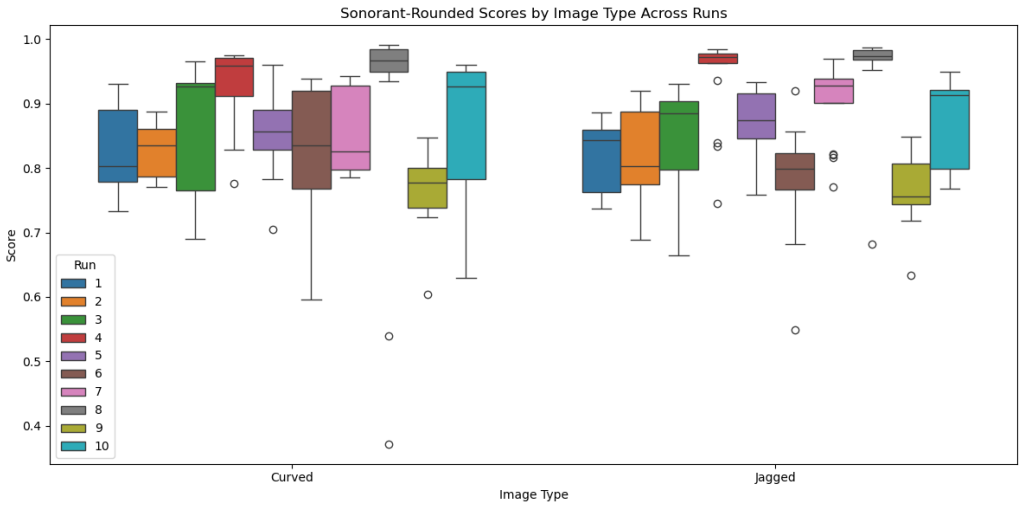

Figure 19: **Molmo**: Average scores for Sonorant-Rounded words across all 10 iterations.

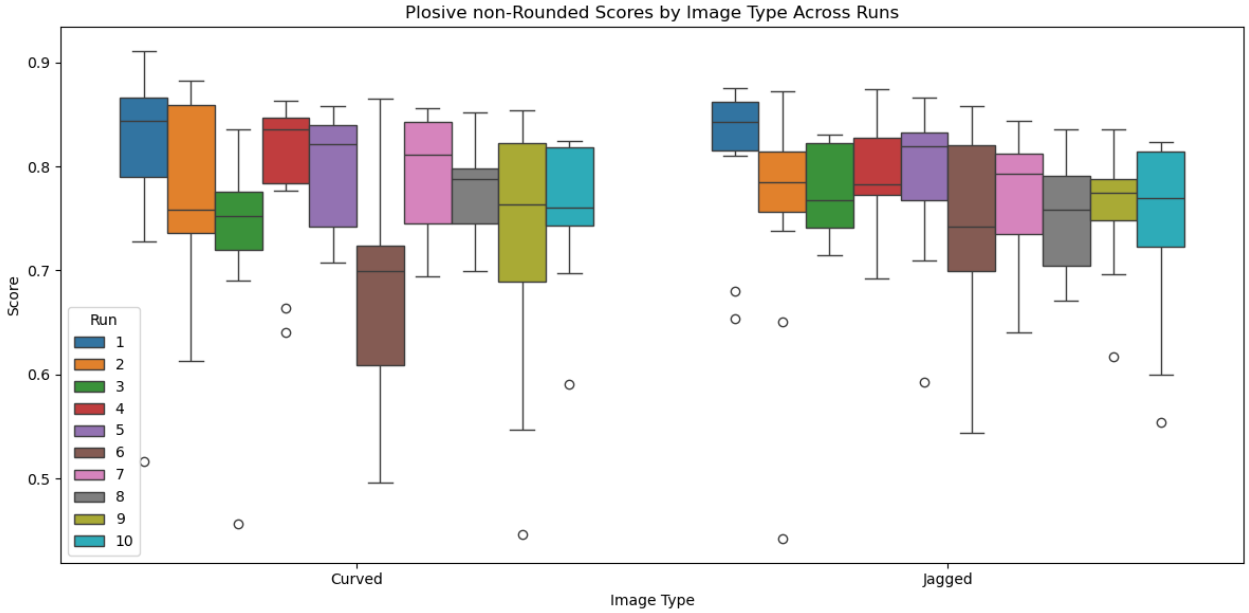

Figure 20: **LLaMA**: Average scores for Plosive-non-Rounded words across all 3 iterations.

Figures 20 and 21 display the difference in accompanying confidence levels regarding plosive non-rounded words on both curved and jagged images, returned by both models.

In figure 20, the average scores for P-NR words is highly inconsistent, showing similar values across the two image types. However, since the scores on jagged images are more balanced and show less variability, we may speak of higher confidence in the model's decision-making with regards to a kiki-like association between plosive non-rounded words and jagged imagery.

In contrast, figure 21 shows significantly more variability, suggesting that Molmo has not successfully found any kiki-like association.

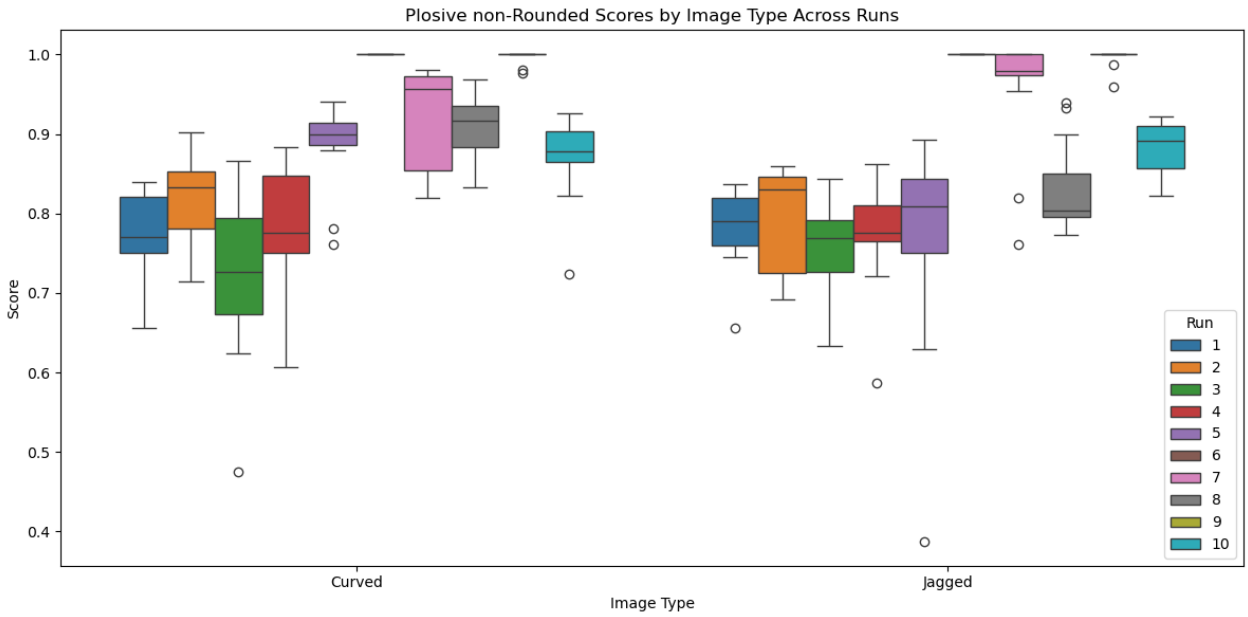

Figure 21: **Molmo**: Average scores for Plosive-non-Rounded words across all 3 iterations.

4.2.4 Significance Analysis

Type	Mean Curved	Mean Jagged	t-value	p-value
S-R	0.801	0.807	-0.559	0.579
P-NR	0.773	0.767	0.782	0.434

Table 5: Statistical Significance test (Welch's t-test) results on LLaMA.

To verify whether these findings are statistically significant, we have conducted a Welch's t-test. The reason for this is that in neither results the variance was equally distributed, therefore the Welch's t-test is a better fit than the Student's t-test. The results in tables 5 and 6 suggest that instead of a congruency pattern befitting the Bouba-Kiki effect, we speak of a pattern that may solely reflect a preference based on words. The p-values are too low to signify statistical importance.

Type	Mean Curved	Mean Jagged	t-value	p-value
S-R	0.8525	0.859	-0.680	0.497
P-NR	0.855	0.869	-1.241	0.215

Table 6: Statistical significance test (Welch's t-test) results on Molmo.

4.3 Attention Pattern Analysis

In this experiment we investigated the visual grounding and referential focus of Molmo in a forced-choice comprehension task. The 4 randomly sampled Bouba-Kiki pseudowords used in this experiment include: ”mahnoo” (bouba), ”lohmah” (bouba), ”teepuh” (kiki), and ”kaytay” (kiki).

4.3.1 Congruent vs Incongruent Shapes

Congruent pairings can contextually be interpreted as an instance of the model pointing towards the correct shape, as demonstrated in figures 22 and 23.

Figure 22: Molmo correctly identifies the curved shape by pointing to what it believes is ”lohmah”.

Figure 23: Molmo correctly identifies the curved shape by pointing to what it believes is ”mahnoo”.

In contrast, an incongruent pairing can be understood as an instance of the model pointing towards the wrong shape, as demonstrated in figure 24. Since ”teepuh” is a kiki-pseudoword, the model should have pointed at the jagged image.

Figure 24: Molmo incorrectly identifies the jagged shape by pointing to what it believes is ”teepuh”.

Similarly to figures 22-24, correct instances of congruent pairings are highlighted in figures 25-27. The model consistently points towards the correct shape, including special instances such as shapes within other shapes as included in figure 25.

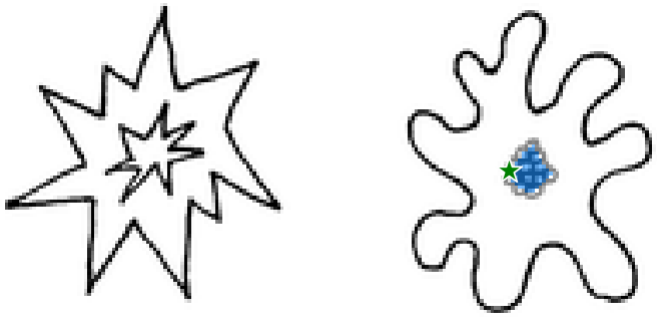

Figure 25: Molmo correctly identifies the curved shape by pointing to what it believes is ”lohmah”.

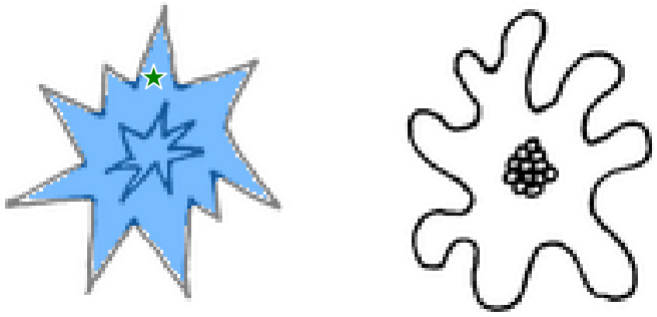

Figure 26: Molmo correctly identifies the jagged shape by pointing to what it believes is ”teepuh”.

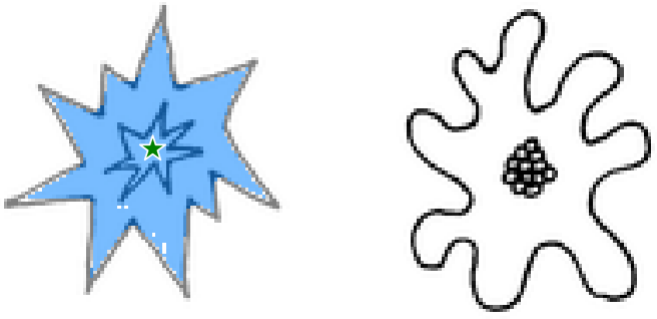

Figure 27: Molmo correctly identifies the jagged shape by pointing to what it believes is ”teepuh”.

Instances of misclassification happening are highlighted in figures 28-30. The shapes included in these figures are highly abstract, therefore likely confusing the model. Nevertheless, these instances reinforce the argument of Molmo not exhibiting the Bouba-Kiki effect.

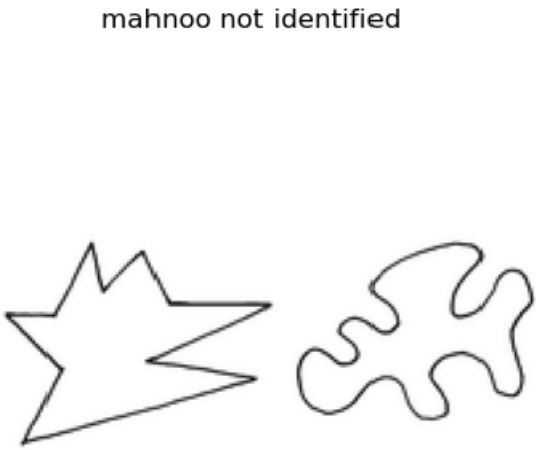

Figure 28: Molmo incorrectly identifies the curved shape by pointing to what it believes is ”mahnoo”.

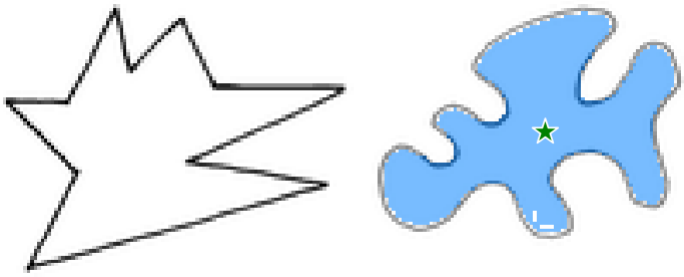

Figure 29: Molmo incorrectly identifies the jagged shape by pointing to what it believes is ”kaytay”.

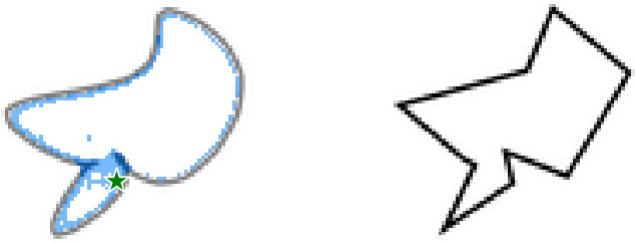

Figure 30: Molmo incorrectly identifies the jagged shape by pointing to what it believes is ”teepuh”.

4.3.2 Congruency Evaluation

Besides the number of trials included in figures 22-24, the remainder of the trials are included in appendix B. The total results of the third experiment are included in table 7.

The results demonstrate that Molmo was able to find more congruent pairings for curved images (approx. 61%) than for jagged images (approx. 47%) in the context of the Bouba-Kiki effect. Please note that special instances such as figure 31 are deemed correct although the blue marking is outside of the shape.

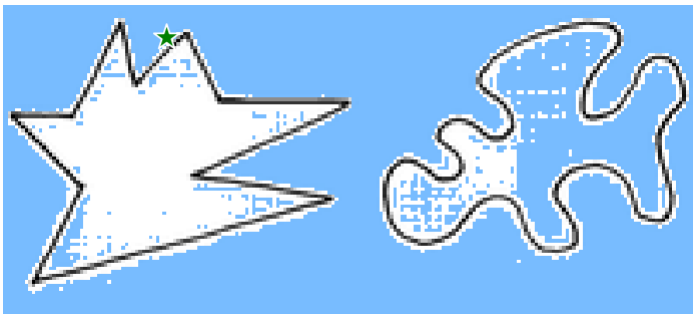

Figure 31: Molmo correctly identifies the jagged shape by pointing to what it believes is ”teepuh”.

Condition	Congruent Cases	Total Trials
Curved Congruency	22	36
Jagged Congruency	17	36

Table 7: Number of congruent cases out of total trials for curved and jagged images.

These results suggest that there is a moderate tendency to pair curved images with sonorant-rounded pseudowords. This aligns somewhat with the Bouba-Kiki hypothesis, but does not deliver overwhelming proof or evidence in favor of it.

Furthermore, since the probability of jagged congruency is below 50%, it suggests that the jagged images were less likely to be matched with plosive non-rounded pseudowords than random guessing would predict.

These arguments suggest that there is weak evidence supporting the Bouba-Kiki effect for curved stimuli, but not for jagged ones. The asymmetry may imply that curved stimuli elicits stronger associations than jagged ones, or that jagged mappings are less consistent or overall more difficult to find.

4.3.3 Statistical Significance Analysis

The results in tables 7 and 8 highlight that the frequency distribution of congruent mappings is not significantly higher than 50%, thereby indicating a lack of statistical significance of the findings gained from this experiment.

Congruency	Proportion	p-value
Curved	0.243	0.6111
Jagged	0.868	0.472

Table 8: Significance analysis binomial test results of Molmo.

5 Summary of Results

This thesis set out to investigate whether cutting-edge vision-language models, LLaMA3.2 and Molmo, exhibit cross-modal associations between visual shape and phonological form, more specifically, the phenomenon famously illustrated by the Bouba-Kiki effect in human cognition. Using LLaMA3.2 and Molmo as test models, we designed and conducted a series of zero-shot experiments to examine model preferences for matching abstract sonorant-rounded and plosive non-rounded pseudowords, respectively matching the idea behind ”bouba” and ”kiki”, to either curved or jagged shapes. Across three experiments, we measured semantic fit through token probabilities, direct classification, and attention patterns.

The first experiment detailed an experimental set-up to perform cross-modal probability analysis by prompting the models to return a score between 0 to 100 on how well an abstract word, sampled from a list of pseudowords, matches a given image. The results of this experiment indicate that both LLaMA3.2 and Molmo have inconsistent congruency patterns, thereby providing weak support in favor of the Bouba-Kiki hypothesis within these models. Moreover, the effect strength measured across these experiments, which denotes sensitivity to shape-sound congruency, was found to be too minor for LLaMA, whereas the results of Molmo gave slightly stronger albeit still relatively weak support, although notably Molmo was found to be more certain overall in its decision-making ability. After comparing the assigned probabilities to curved and jagged images, it was discovered that neither LLaMA or Molmo had consistent congruency patterns over all experiment runs. By analyzing the assigned probabilities per image type in higher detail, it was found that LLaMA had an inconsistent pattern of assigning higher scores to congruent pairings than to incongruent pairings, whereas Molmo had a similar, albeit more balanced range of assigned probabilities, which may benefit from further investigation.

The second experiment detailed an experimental set-up to perform an image-to-text matching task. The aim was to assess whether LLaMA or Molmo show, on average, higher token-wise confidence levels for bouba- and kiki-pseudowords on congruent pairings of curved and jagged images. The experiment was set up in a way that enables the model to ask, for every image, which word from a subset of sonorant-rounded and plosive non-rounded words, matched the best with the image. The extracted results contain both the pseudowords as well as the confidence score of the token. The findings of this experiment show that the average scores returned by both LLaMA and Molmo are not higher on congruent pairings than incongruent pairings. Moreover, the detailed score differences themselves also do not indicate a substantial difference. A comparison of the scores of sonorant-rounded and plosive non-rounded over all iterations highlight that LLaMA assigns nearly identical confidence scores to sonorant-rounded and plosive non-rounded words for both curved and jagged images, with minor variability. In contrast, Molmo includes higher variability in score differences, although still lacking in a pattern that supports the expected results aligning with the Bouba-Kiki effect.

The third experiment detailed an experimental set-up built on `SAM2` to perform attention pattern analysis. More specifically, an experiment to highlight whether Molmo's [4] spatial attention corresponded more closely to a curved or jagged side of a bipartite image, depending on shape-sound congruency. The results of this experiment were congruency accuracies of approximately 61% for curved images and 47% for jagged images. These results indicate that we may speak of a moderate tendency to pair curved images with sonorant-rounded pseudowords. All in all, the findings do not incur strong evidence supporting the Bouba-Kiki effect for Bouba-curved and Kiki-jagged associations.

[4] *LLaMA support for this experiment does not exist yet.

Taken together, these 3 experiments suggest that while LLaMA3.2 and Molmo exhibit some degree of sensitivity to shape-sound congruency, the evidence remains limited and highly inconsistent. LLaMA's behavior appears closer to random across conditions, while Molmo demonstrates slightly more stable, though still inconclusive, preferences. Notably, both models show a weak inclination to associate curved stimuli with sonorant-rounded pseudowords, aligning partially with the Bouba-Kiki hypothesis. However, this tendency does not extend reliably to jagged images or plosive-like sounds. Therefore, while vision-language models may encode rudimentary cross-model associations by random chance, since the results highlight that the models do not consistently exhibit cognitive-like shape-sound congruency, we do not have any evidence in favor of the Bouba-Kiki hypothesis in these models.

6 Discussion

The results of this thesis provide an initial yet nuanced perspective on the capacity of large vision-language models to form cross-modal associations between visual and textual features. Drawing inspiration from the famous Bouba-Kiki effect, this work examined whether two advanced, cutting-edge models, LLaMA3.2 and Molmo, show similar tendencies when tasked with associating pseudowords to either curved or jagged shapes. While some patterns emerged that may resemble human-like congruency, especially in the case of curved shapes paired with sonorant-rounded pseudowords, the overall findings suggest only weak and inconsistent alignment with the expected shape-sound mappings. This invites a deeper reflection on the nature of such associations within deep, complex neural and transformer-based models, and on whether cognitive-like behavior can truly emerge from training objectives that do not explicitly involve such embodied and grounded relationships.

6.1 Interpretation of Congruency Patterns

Across the 3 experiments, when present, congruency effects were subtle and often inconsistent. In the cross-modal probability assignment task, neither LLaMA3.2 nor Molmo showed robust alignment between curved shapes and sonorant-rounded pseudowords or jagged shapes and plosive non-rounded pseudowords across all runs. Although Molmo produced slightly higher probabilities for congruent pairs in some cases, the effect was small and unstable, thereby suggesting that such associations may be partially represented but not systematically encoded.

The image-to-text matching experiment provided further nuance. While both models occasionally assigned marginally higher confidence scores to congruent word-shape pairings, these differences were not statistically strong and varied frequently across iterations. Notably, LLaMA's output was almost flat in terms of score differentiation, whereas Molmo demonstrated greater variability, potentially hinting at model-specific mechanisms or the PixMo dataset influencing confidence assignment. However, in both cases, the signal remained too insignificant to conclude that either model has learned or discovered a robust cross-modal mapping comparable to human intuition.

The third experiment, focusing on spatial attention using a bipartite shape template, offered a more visually grounded perspective into word-shape associations. While a modest tendency was found in Molmo to attend more strongly to the curved region of an image when prompted with a congruent pseudoword, this effect did not generalize to jagged regions or across model conditions. Therefore, attention spans, while visually compelling, do not provide strong corroborative evidence for shape-sound congruency.

In summary, these results suggest that while large vision-language model may incidentally reflect aspects of expected Bouba-Kiki alignment, particularly in curved-shape contexts, these associations are weak, highly-prompt sensitive, and are too inconsistent to indicate genuine internalization of the Bouba-Kiki effect.

6.2 Comparing LLaMA and Molmo

A central aim of this thesis was to compare a general-purpose vision-language model (LLaMA3.2) with a more specialized and open model (Molmo) to investigate whether different architectures or training objectives influence the emergence of cross-modal associations. The results across all 3 experiments suggest notable differences in behavior and interpretability between the models.

LLaMA3.2, despite its strong general language and visual capabilities, demonstrated minimal sensitivity to shape-sound congruency. Its responses in both the probability estimation and classification tasks were largely inconsistent, often assigning irregular scores to congruent and incongruent pairings, besides some exceptions. In some cases, it even displayed reversed patterns by favoring incongruency over congruency. This lack of differentiation indicates that LLaMA's internal representations may not prioritize or encode subtle symbolic mappings such as the underlying Bouba-Kiki effect. Alternatively, it could also reflect that LLaMA, optimized for broad instruction-following and generative coherence, lacks the inductive biases necessary to spontaneously draw connections between sound and shape.

Molmo, in contrast, produced slightly more consistent behavior. It showed marginal but measurable differences between congruent and incongruent pairings in both score-based and attention-based analyses. Its confidence scores fluctuated more dynamically in response to input differences, and its attention maps revealed some alignment with visually relevant regions, specifically for curved-soft sound-shape congruency. While not strong enough to conclude that Molmo has learned a cognitive-like mapping, these patterns suggest a more fine-grained perceptual sensitivity, potentially due to differences in architecture or training methods.

Ultimately, while neither model convincingly encodes robust cross-modal congruency patterns, Molmo shows signs of a closer fit. This highlights the importance of model design in studying emergent behaviors and suggests that model comparisons may offer valuable insights into their nature and limitations regarding perceptual reasoning in artificial systems such as VLMs. These findings highlight both the promise and the current limitations of using multimodal foundation models as cognitive models of perception.

6.3 Implications for Cognitive Modeling

The Bouba-Kiki effect has long been interpreted as evidence of innate or universal cross-modal associations in human cognition by linking auditory and visual modalities in a way that appears deeply embedded and intuitive, even in the preverbal stage of human development. One of the motivations for this thesis was to explore whether modern, cutting-edge VLMs, trained at scale and without explicit symbolic grounding, would naturally exhibit similar capabilities. If such models were to show reliable congruency patterns, it could provide support for the hypothesis that cross-modal alignment can emerge from exposure or statistical patterns within both textual and visual data, even without prior training, thereby including them as an emergent behavior.

The experimental findings presented here, however, suggest that VLMs do not reliably encode such associations, as of now. Both LLaMA and Molmo showed weak and inconsistent alignment with shape-sound mappings, besides some marginal exceptions. This raises important questions about the extent to which such models offer utilitarian benefits for the analyzing and understanding of human cognitive processes. While VLMs demonstrated impressive generalization and multimodal reasoning capabilities, they lack the fundamental understanding that underpin human-like perception and intuition.

That said, the partial effects observed, particularly by Molmo's slight tendency to score curved, sonorant-rounded associations higher than its incongruent counterpart, suggest that large models may be capable of approximating cognitive intuition under certain conditions. This opens up potential future research paths, particularly in identifying what training conditions, architectures, or constraints would be necessary for models to internalize more human-like multimodal correspondences. In this sense, vision-language models offer both challenges and an opportunities. They are not yet fully functional mini-humans, but their behavior provides a

window into the kinds of structure that might support cognitively plausible reasoning, potentially resulting in a system that can grow to be on par with humans.

6.4 Challenges in Measuring Cross-Modal Semantics

Evaluating cross-modal semantics in foundation models presents several methodological challenges. Unlike traditional classification tasks with discrete accuracy metrics, cross-modal association relies on subtle, often ambiguous mappings between modalities. Contextually, it relies on abstract pseudowords and geometric shapes. These abstract pseudowords and geometric shapes may all carry their own, separate understandings and could create difficulty both in designing effective prompts and in interpreting model outputs reliably.

One major challenge lies in prompt sensitivity. The results of this thesis show that even slight changes in wording, format, or sampling parameters, may significantly influence a model's output. This sensitivity makes it difficult to disentangle genuine internal associations from superficial prompt-following behavior. For instance, a model could generate different scores for the image-pseudoword pair depending on how the prompt is phrased or how many labels are provided as options. This complicates the task of evaluating whether the model's behavior reflects learned multimodal concepts or merely linguistic compliance.

Additionally, the inherently abstract nature of the Bouba-Kiki task, which relies on pseudowords with no semantic referents, limits the influence of pretraining data and requires models to generalize in a cognitively plausible way. However, large language models are often optimized for pragmatic coherence rather than symbolic generalization, and may simply lack the architectural or representational foundation to form such associations in a consistent manner.

Another limitation is interpretability. While token-level confidence scores and attention heatmaps provide valuable information for internal model behavior, they remain indirect. A high token probability may reflect generation fluency rather than a genuine association between modalities. Similarly, attention maps may not directly correspond to the model's decision-making and its motivation. These metrics are useful but imperfect, and future methods may need to combine behavioral probing with more advanced techniques.

7 Limitations

While this thesis presents a framework for probing cross-modal associations in vision-language models, several limitations constrain the generalizability and interpretability of its findings.

First, the experiments rely heavily on zero-shot prompting, which can be extremely sensitive to phrasing, tokenization, and sampling parameters. Despite efforts to standardize prompts and use diverse word sets, minor variations in temperature, top-p, or word order may have influenced the model outputs in unintended ways. This makes it challenging to isolate the effects of true semantic congruency from model-specific generation artifacts. Moreover, forcing the model to restrict outputs to numeric values or single label types may have obstructed the models from truly finding the connection between sound and shape, considering a scenario where the model would consistently associate soft words with curved shapes, while in the experiment this output was limited to a single word.

Second, the use of pseudowords, while appropriate for studying Bouba-Kiki effects, introduce ambiguity. Unlike real-world concepts, pseudowords lack proper grounding in model pretraining data, meaning any emergent associations are either abstract or incidental. While this makes the task cognitively interesting, for a model that follows its architecture this may not carry any human-like intuitive meaning, thereby making the findings limiting.

Finally, the sample size and scope of this study were limited by computational resources. The number of images, pseudowords, and experimental repetitions, while sufficient for explanatory analysis, do not reach the scale required for definite statistical claims for model behavior. This limits the strength of conclusions, particularly when effect sizes are small or vary across different runs.

8 Future Work

Several directions for future research can help refine and extend the importance of the findings of this thesis.

First, future studies could scale the size of the dataset and include more diverse image stimuli and a broader range of pseudowords. Using carefully balanced sets of real and artificial labels could explore whether models encode structural features or simply exploit prompt regularities.

Second, follow-up experiments could explore fine-tuning or in-context learning to see whether models can be taught the shape-sound connection described in the Bouba-Kiki hypothesis. Comparing zero-shot, to few-shot, to many-shot learning, could draw further attention to the models' capacity for internalizing symbolic mappings, if given proper guidance.

Third, further investigation into model architecture is warranted. The contrast between LLaMA3.2 and Molmo suggests that training objectives could influence cross-modal alignment. By increasing the scope of this research to additional cutting-edge models, it could increase understanding of how training scale, data modality imbalances, or structural components may affect sound-shape connections.

Fourth, as mentioned in the first experiment, in general, Molmo assigned higher probability scores to both congruent and incongruent pairings than LLaMA, implying that it is less conservative in assigning scores and thereby more confident in its decision making. By changing the methodology and the test data, this thesis may be expanded on by researching what components or underlying concepts determine this discrepancy between LLaMA and Molmo.

Finally, integrating human behavioral baselines, for example by comparing model outputs with human judgments of pseudoword-shape congruency, would offer an additional benchmark for evaluating model cognition and alignment. This may help to clarify whether models are computationally clever or cognitively aligned.

References

[Alper and Averbuch-Elor, 2023] Alper, M. and Averbuch-Elor, H. (2023). Kiki or bouba? sound symbolism in vision-and-language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc.

[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). 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, page 610–623, New York, NY, USA. Association for Computing Machinery.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Deitke et al., 2024] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.-H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N. A., Hajishirzi, H., Girshick, R., Farhadi, A., and Kembhavi, A. (2024). Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Fort et al., 2018] Fort, M., Lammertink, I., Peperkamp, S., Guevara-Rukoz, A., Fikkert, P., and Tsuji, S. (2018). Symbouki: a meta-analysis on the emergence of sound symbolism in early language acquisition. *Developmental Science*, 21(5):e12659.

[Fort and Schwartz, 2022] Fort, M. and Schwartz, J.-L. (2022). Resolving the bouba-kiki effect enigma by rooting iconic sound symbolism in physical properties of round and spiky objects. *Scientific Reports*, 12(1):19172.

[Grattafiori et al., 2024] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe,

J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R.,

Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models.

[Imai et al., 2008] Imai, M., Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65.

[Islam et al., 2023] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., and Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks.

[Jain and Wallace, 2019] Jain, S. and Wallace, B. C. (2019). Attention is not explanation.

[Köhler, 1929] Köhler, W. (1929). *Gestalt Psychology*. Horace Liveright, New York.

[Köhler, 1947] Köhler, W. (1947). *Gestalt Psychology. (2nd ed.)*. Horace Liveright, New York.

[Li et al., 2023] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

[Loakman et al., 2024] Loakman, T., Li, Y., and Lin, C. (2024). With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2849–2867, Miami, Florida, USA. Association for Computational Linguistics.

[Marklová et al., 2025] Marklová, A., Milička, J., Ryvkin, L., Ľudmila Lacková Bennet, and Kormaníková, L. (2025). Iconicity in large language models.

[Marks, 1974] Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, 87(1/2):173–188.

[Maurer et al., 2006] Maurer, D., Pathman, T., and Mondloch, C. J. (2006). The shape of boubas: sound–shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–322.

[Meta, 2024] Meta (2024). Introducing llama 3.2: Vision and multimodal capabilities for edge and mobile devices. Accessed: 2025-02-13.

[Nielsen and Rendall, 2012] Nielsen, A. and Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4(2):115–125.

[Ozturk et al., 2013] Ozturk, O., Krehm, M., and Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114(2):173–186.

[Peiffer-Smadja and Cohen, 2019] Peiffer-Smadja, N. and Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *NeuroImage*, 186:679–689.

[Perlman et al., 2015] Perlman, M., Dale, R., and Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, 2(8):150152.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

[Ramachandran and Hubbard, 2001] Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Shahrasbi, 2024] Shahrasbi, K. (2024). Exploring the bouba-kiki effect: Cross-modal associations in vision-and-language models. Master's thesis, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands. Specialisation: Artificial Intelligence.

[Strubell et al., 2020] Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Verhoef et al., 2024] Verhoef, T., Shahrasbi, K., and Kouwenhoven, T. (2024). What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213.

[Webson and Pavlick, 2022] Webson, A. and Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

[Wei et al., 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models.

[Yatskar et al., 2016] Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ćwiek et al., 2022] Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, , Ünal Logacev, , Zeller, J., Perlman, M., and Winter, B. (2022). The ¡i¿bouba/kiki¡/i¿ effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200390.

9 Appendices

9.1 Appendix A

Congruent Stimuli (S-R) for Curved Shapes

looloo	nooloo	mooloo
loolah	noolah	moolah
looloh	nooloh	mooloh
loonoo	noonoo	moonoo
loonah	noonah	moonah
loonoh	noonoh	moonoh
loomoo	noomoo	moomoo
loomah	noomah	moomah
loomoh	noomoh	moomoh
lahloo	nahloo	mahloo
lahlah	nahlah	mahlah
lahloh	nahloh	mahloh
lahnoo	nahnoo	mahnoo
lahnah	nahnah	mahnah
lahnoh	nahnoh	mahnoh
lahmoo	nahmoo	mahmoo
lahmah	nahmah	mahmah
lahmoh	nahmoh	mahmoh
lohloo	nohloo	mohloo
lohlah	nohlah	mohlah
lohloh	nohloh	mohloh
lohnoo	nohnoo	mohnoo
lohnah	nohnah	mohnah
lohnoh	nohnoh	mohnoh
lohmoo	nohmoo	mohmoo
lohmah	nohmah	mohmah
lohmoh	nohmoh	mohmoh

Congruent Stimuli (P-NR) for Jagged Shapes

teetee	tuhtay	taykuh
teetuh	tuhkee	taykay
teetay	tuhkuh	taypee
teekee	tuhkay	taypuh
teekuh	tuhpee	taypay
teekay	tuhpuh	keetee
teepee	tuhpay	keetuh
teepuh	taytee	keetay
teepay	taytuh	keekee
tuhtee	taytay	keekuh
tuhtuh	taykee	keekay

keepee
keepuh
keepay
kuhtee
kuhtuh
kuhtay
kuhkee
kuhkuh
kuhkay
kuhpee
kuhpuh
kuhpay
kaytee
kaytuh
kaytay
kaykee
kaykuh
kaykay
kaypee
kaypuh
kaypay
peetee
peetuh
peetay
peekee
peekuh
peekay
peepee
peepuh
peepay
puhtee
puhtuh
puhtay
puhkee
puhkuh
puhkay
puhpee
puhpuh
puhpay
paytee
paytuh
paytay
paykee
paykuh
paykay
paypee
paypuh
paypay

9.2 Appendix B

Bouba-pseudowords: mahnoo, lohmah
Kiki-pseudowords: teepuh, kaytay
Blue markings highlight the focus points of Molmo.

Block 1

Figure 32: *
”point to mahnoo”

Figure 33: *
”point to lohmah”

Figure 34: *
”point to teepuh”

Figure 35: *
”point to kaytay”

Figure 36: *
”point to mahnoo”

Figure 37: *
”point to lohmah”

Figure 38: *
”point to teepuh”

Figure 39: *
”point to kaytay”

Block 2

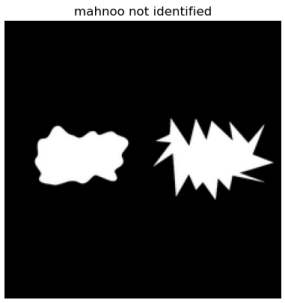

Figure 40: * ”point to mahnoo”

Figure 41: * ”point to lohmah”

Figure 42: * ”point to teepuh”

Figure 43: * ”point to kaytay”

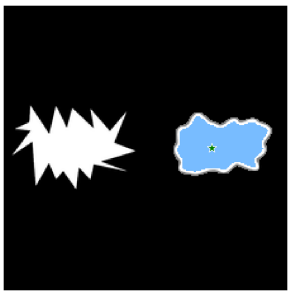

Figure 44: * ”point to mahnoo”

Figure 45: * ”point to lohmah”

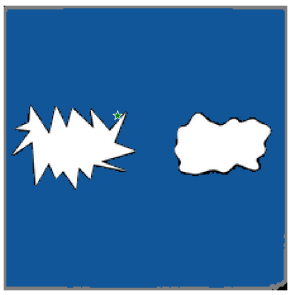

Figure 46: * ”point to teepuh”

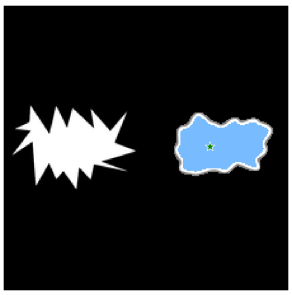

Figure 47: * ”point to kaytay”

Block 3

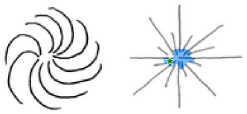

Figure 48: *
”point to mahnoo”

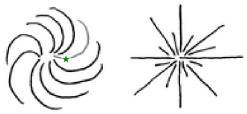

Figure 49: *
”point to lohmah”

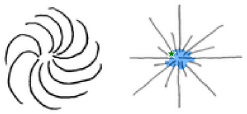

Figure 50: *
”point to teepuh”

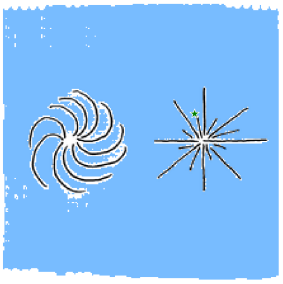

Figure 51: *
”point to kaytay”

Figure 52: *
”point to mahnoo”

Figure 53: *
”point to lohmah”

Figure 54: *
”point to teepuh”

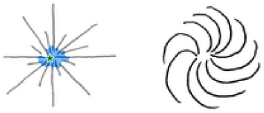

Figure 55: *
”point to kaytay”

Block 4

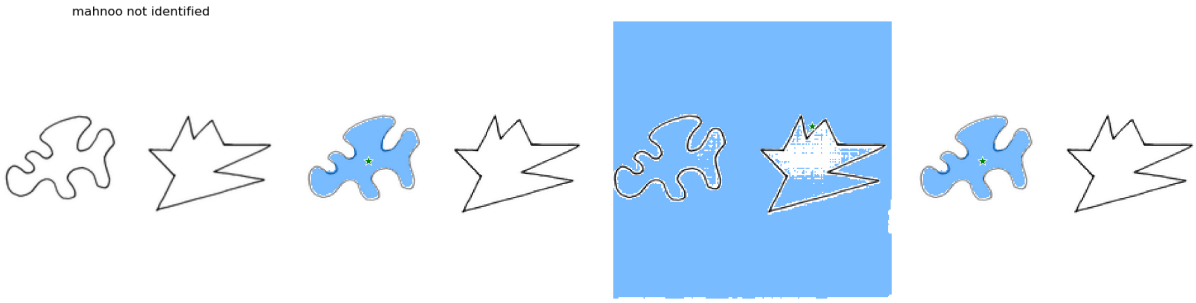

Figure 56: * ”point to mahnoo”

Figure 57: * ”point to lohmah”

Figure 58: * ”point to teepuh”

Figure 59: * ”point to kaytay”

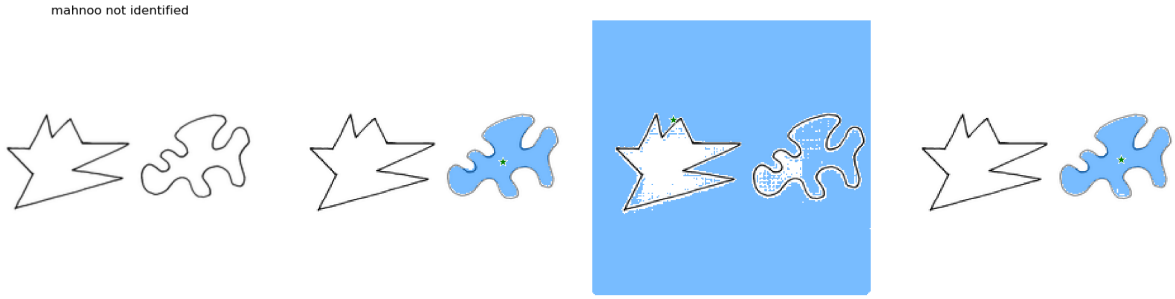

Figure 60: * ”point to mahnoo”

Figure 61: * ”point to lohmah”

Figure 62: * ”point to teepuh”

Figure 63: * ”point to kaytay”

Block 5

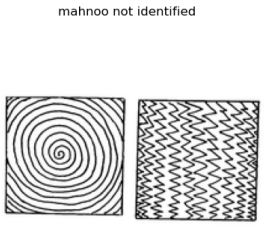

Figure 64: * ”point to mahnoo”

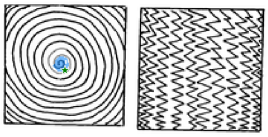

Figure 65: * ”point to lohmah”

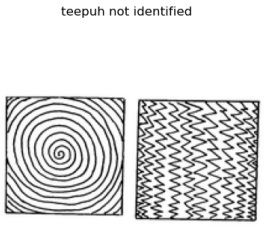

Figure 66: * ”point to teepuh”

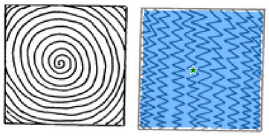

Figure 67: * ”point to kaytay”

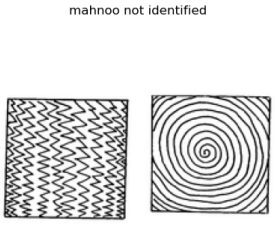

Figure 68: * ”point to mahnoo”

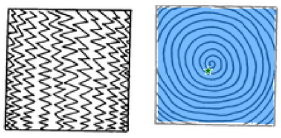

Figure 69: * ”point to lohmah”

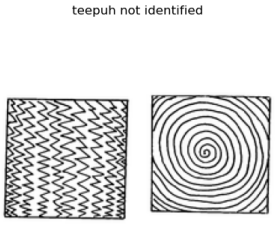

Figure 70: * ”point to teepuh”

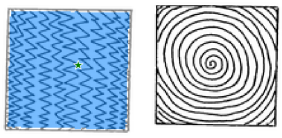

Figure 71: * ”point to kaytay”

Block 6

mahnoo not identified

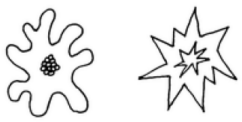

Figure 72: *
”point to mahnoo”

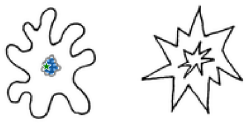

Figure 73: *
”point to lohmah”

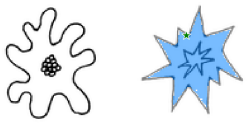

Figure 74: *
”point to teepuh”

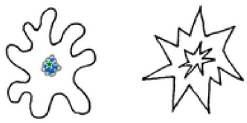

Figure 75: *
”point to kaytay”

mahnoo not identified

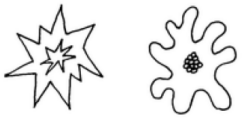

Figure 76: *
”point to mahnoo”

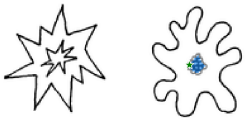

Figure 77: *
”point to lohmah”

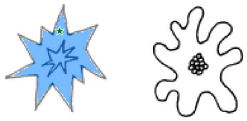

Figure 78: *
”point to teepuh”

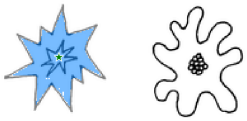

Figure 79: *
”point to kaytay”

Block 7

mahnoo not identified

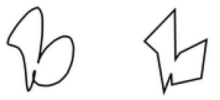

Figure 80: *
”point to mahnoo”

Figure 81: *
”point to lohmah”

Figure 82: *
”point to teepuh”

Figure 83: *
”point to kaytay”

mahnoo not identified

Figure 84: *
”point to mahnoo”

Figure 85: *
”point to lohmah”

Figure 86: *
”point to teepuh”

Figure 87: *
”point to kaytay”

Block 8

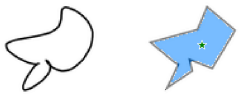

Figure 88: * ”point to mahnoo”

Figure 89: * ”point to lohmah”

Figure 90: * ”point to teepuh”

Figure 91: * ”point to kaytay”

Figure 92: * ”point to mahnoo”

Figure 93: * ”point to lohmah”

Figure 94: * ”point to teepuh”

Figure 95: * ”point to kaytay”

Block 9

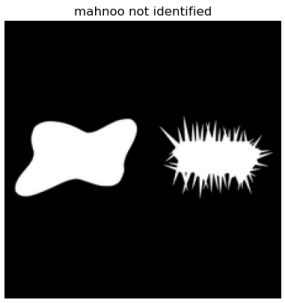

Figure 96: * ”point to mahnoo”

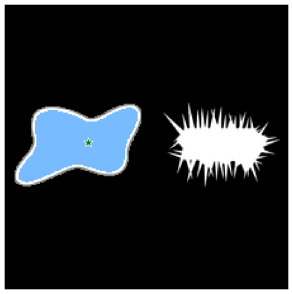

Figure 97: * ”point to lohmah”

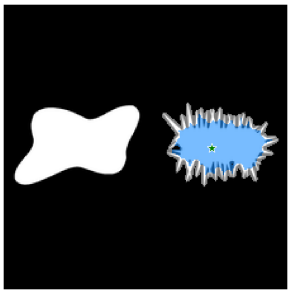

Figure 98: * ”point to teepuh”

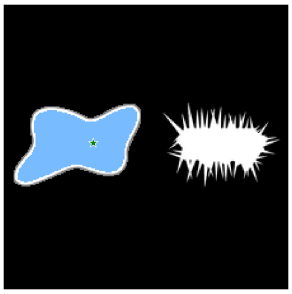

Figure 99: * ”point to kaytay”

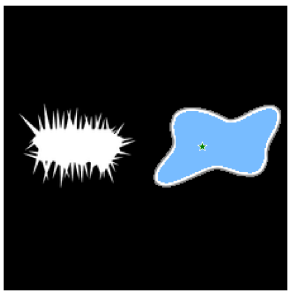

Figure 100: * ”point to mahnoo”

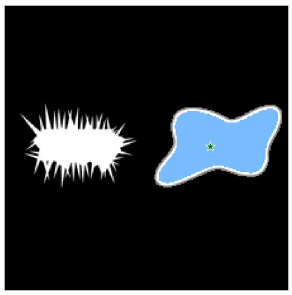

Figure 101: * ”point to lohmah”

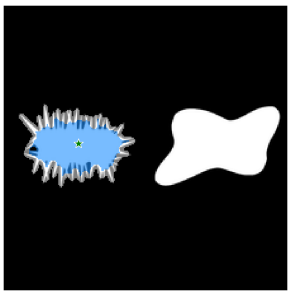

Figure 102: * ”point to teepuh”

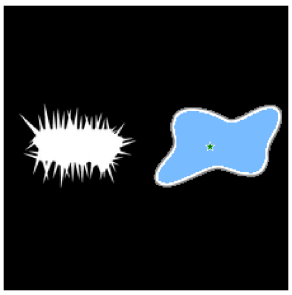

Figure 103: * ”point to kaytay”

9.3 Appendix C

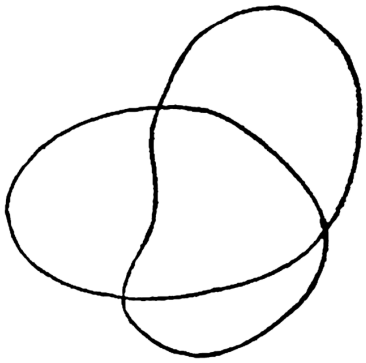

Figure 104: *
1-Kohler curved-p1.png

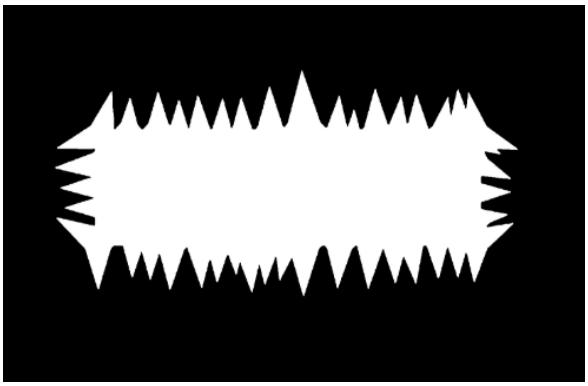

Figure 105: *
10-Westbury4 jagged-p5.png

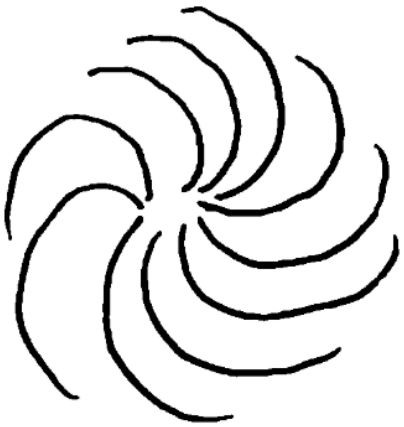

Figure 106: *
11-Maurer1 curved-p6.png

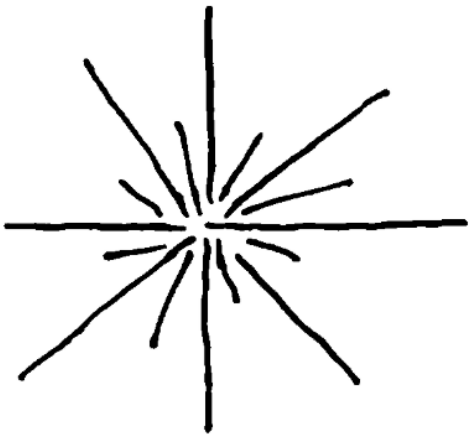

Figure 107: *
12-Maurer1 jagged-p6.png

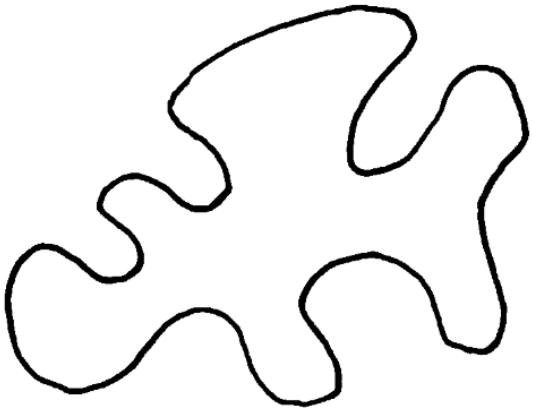

Figure 108: *
13-Maurer2 curved-p7.png

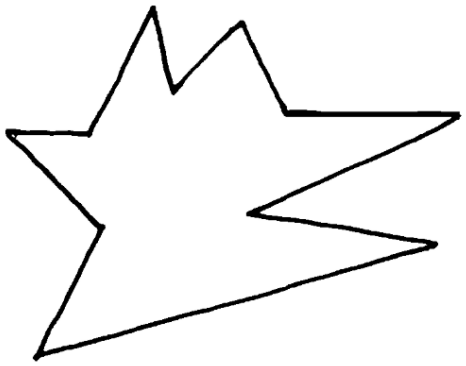

Figure 109: *
14-Maurer2 jagged-p7.png

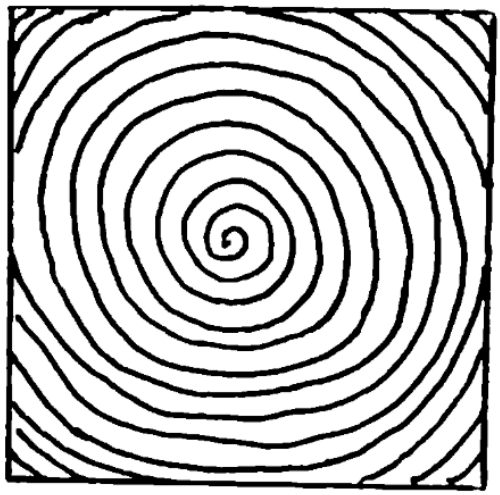

Figure 110: *
15-Maurer3 curved-p8.png

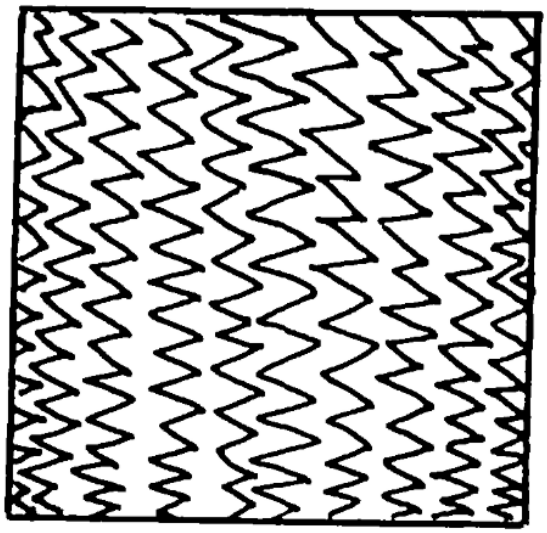

Figure 111: *
16-Maurer3 jagged-p8.png

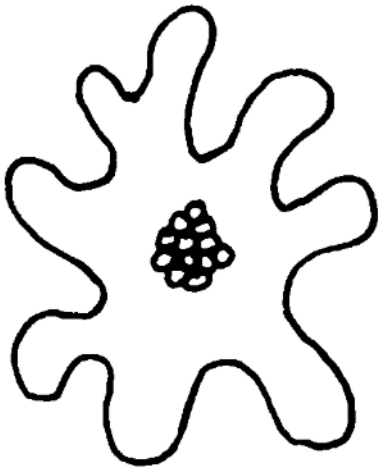

Figure 112: *
17-Maurer4 curved-p9.png

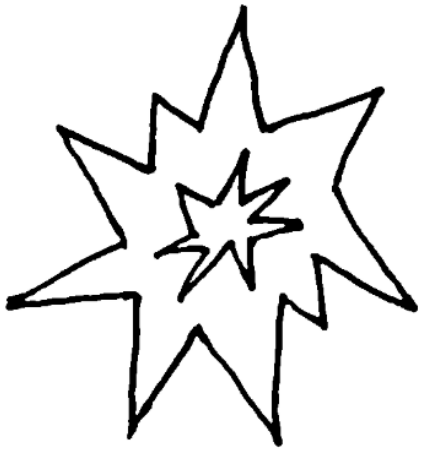

Figure 113: *
18-Maurer4 jagged-p9.png

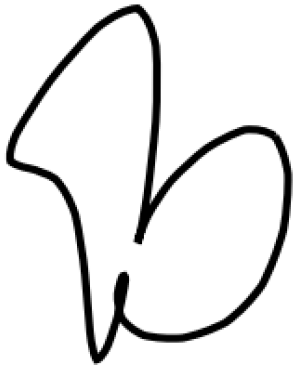

Figure 114: *
19-curved-p10.png

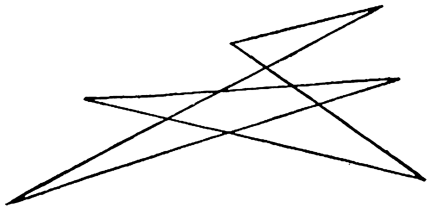

Figure 115: *
2-Kohler jagged-p1.png

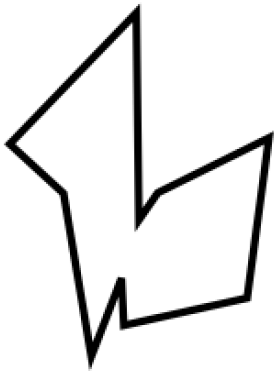

Figure 116: *
20-jagged-p10.png

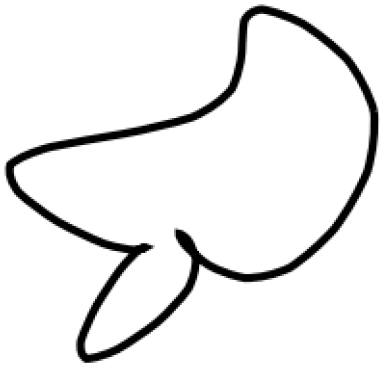

Figure 117: *
21-curved-p11.png

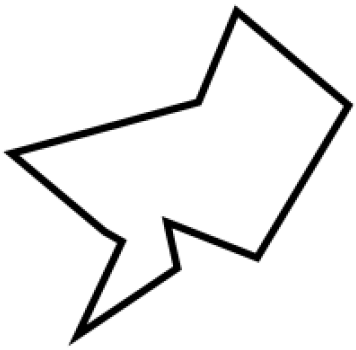

Figure 118: *
22-jagged-p11.png

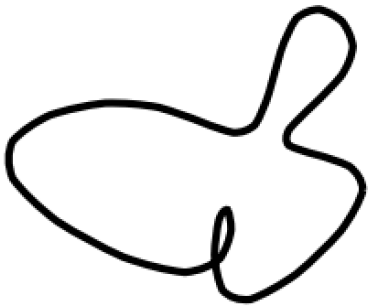

Figure 119: *
23-curved-p12.png

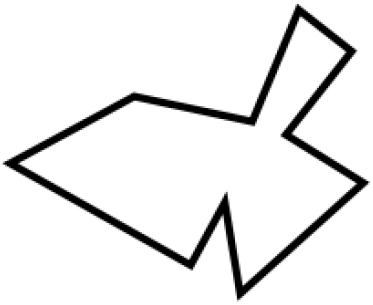

Figure 120: *
24-jagged-p12.png

Figure 121: *
25-curved-p13.png

Figure 122: *
26-jagged-p13.png

Figure 123: *
27-curved-p14.png

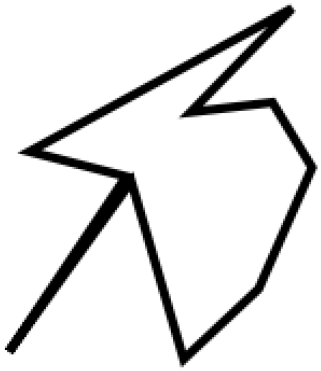

Figure 124: *
28-jagged-p14.png

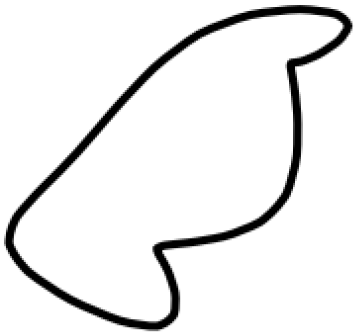

Figure 125: *
29-curved-p15.png

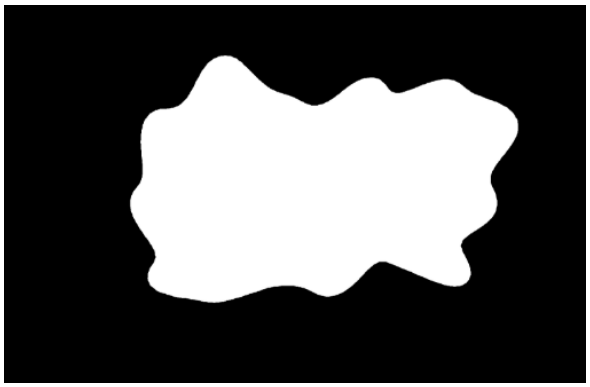

Figure 126: *
3-Westbury1 curved-p2.png

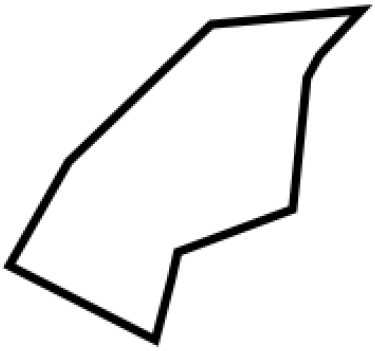

Figure 127: *
30-jagged-p15.png

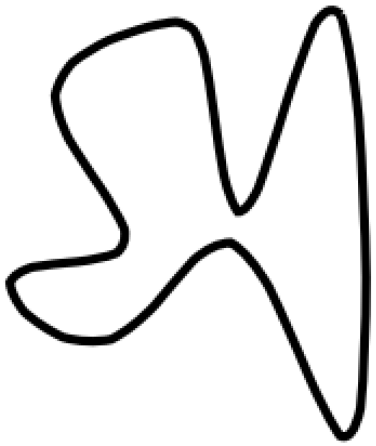

Figure 128: *
31-curved-p16.png

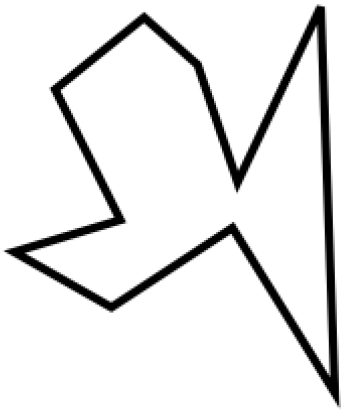

Figure 129: *
32-jagged-p16.png

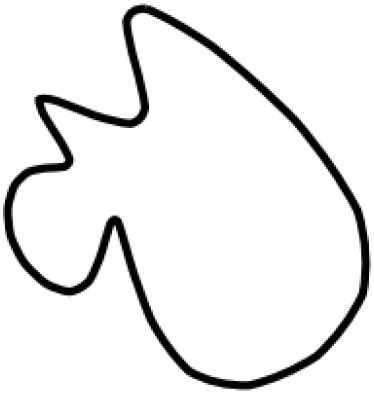

Figure 130: *
33-curved-p17.png

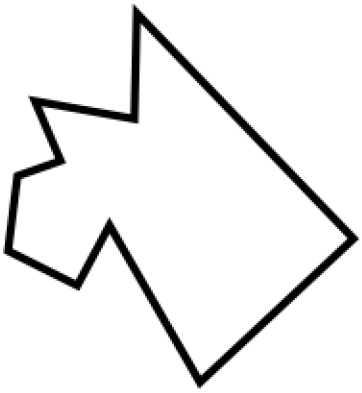

Figure 131: *
34-jagged-p17.png

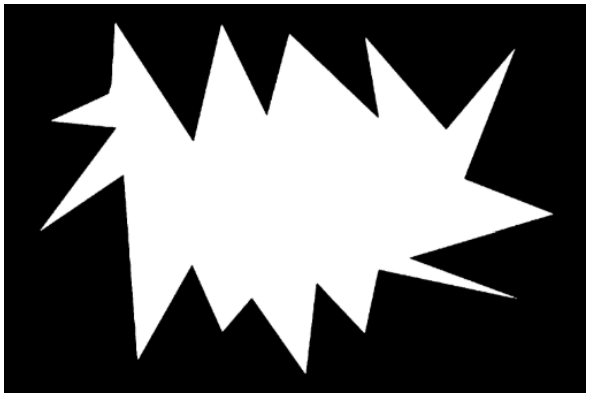

Figure 132: *
4-Westbury1 jagged-p2.png

Figure 133: *
5-Westbury2 curved-p3.png

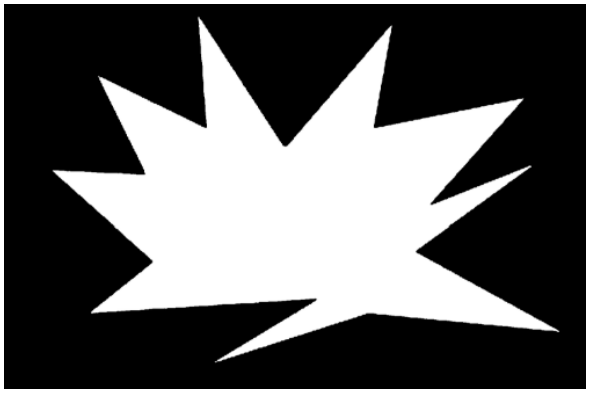

Figure 134: *
6-Westbury2 jagged-p3.png

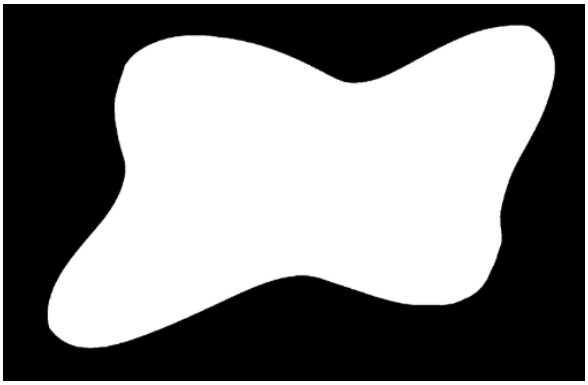

Figure 135: *
7-Westbury3 curved-p4.png

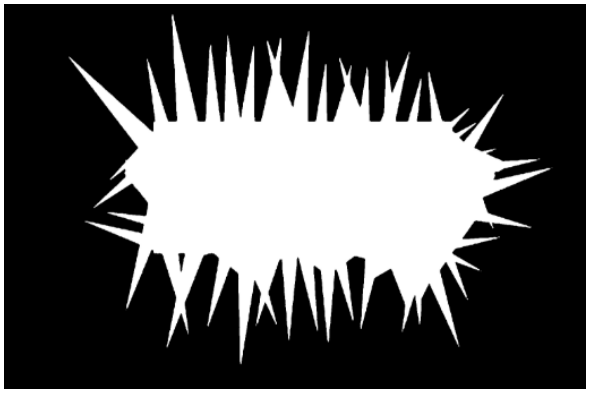

Figure 136: *
8-Westbury3 jagged-p4.png

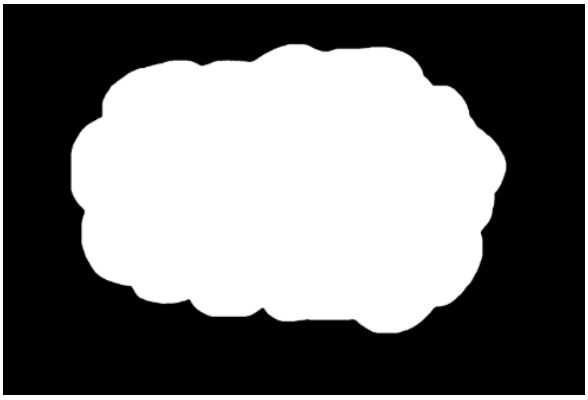

Figure 137: *
9-Westbury4 curved-p5.png