UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Kim Lilii Tamm

Modelling Masculinity Ideals Using Machine Learning: Opinion Leader Discourse Analysis Through Debate

Bachelor's Thesis (9 ECTS)

Supervisor:
Krista Liin, MSc

Tartu 2026

Modelling Masculinity Ideals Using Machine Learning: Opinion Leader Discourse Analysis Through Debate

Abstract:

This thesis designs, implements, and validates a computational pipeline that combines retrieval-augmented generation, multi-agent debate, and the SCALE collaborative annotation framework to analyse the masculinity discourse of YouTube thought leaders.

A corpus of 137 unique YouTube videos from 14 English-language speakers — representing four discursive orientations in contemporary masculinity debate (radical "Red Pill" ideology, discipline and self-optimisation, psychology and relationships, and social criticism) — was created and stored in ChromaDB as retrieval-ready chunks. The SCALE framework was adapted to classify transcript material into three rhetorical categories: Core Dogma, Provocation, and Rhetorical Evidence. The annotated subset is integrated with the full corpus through a two-tier retrieval architecture that supports category-aware boosting without restricting retrieval to the annotated material. The debate system implements a fixed three-round structure mapped to the three rhetorical categories, with a four-criterion faithfulness filter and an evidence-aware moderator that injects cross-speaker counter-evidence between rounds.

The pipeline was evaluated for reliability of the SCALE annotation, faithfulness of the generated debate responses, and correlation between stated positions and observed convergence across the 91 speaker pairs.

Keywords: masculinity discourse, large language models, retrieval-augmented generation, multi-agent debate, SCALE annotation, computational linguistics, YouTube

CERCS: P176 Artificial intelligence

Maskuliinsuse ideaalide modelleerimine masinõppe abil: arvamusliidrite diskursuse analüüs debati kaudu

Lühikokkuvõte:

Käesolev bakalaureusetöö kavandab, realiseerib ja valideerib arvutusliku töövoo, mis ühendab otsingutel põhineva genereerimise (RAG), mitme agendi debati ja kollaboratiivse märgendamise raamistiku SCALE, et analüüsida YouTube'i arvamusliidrite maskuliinsuse diskursust.

Loodi 137 transkribeeritud YouTube'i videost koosnev korpus 14 ingliskeelselt kõnelejalt, kes esindavad kaasaegse maskuliinsuse arutelu nelja diskursiivset orientatsiooni — radikaalset "Red Pill" ideoloogiat, distsipliini ja eneseoptimeerimist, psühholoogiat ja suhteid ning sotsiaalkriitikat — ning see salvestati ChromaDB-sse otsinguvalmis tekstisegmentidena. SCALE raamistik kohandati transkriptimaterjali liigitamiseks kolme retoorilisse kategooriasse: Tuumdogma, Provokatsioon ja Retoorilised tõendid. Märgendatud alamkogum integreeriti kogu korpusega kahetasandilise otsinguarhitektuuri kaudu, mis võimaldab kategooriapõhist eelistamist, ilma et otsing piirduks märgendatud materjaliga. Debatisüsteem koosneb kolmest voorust, kus iga voor on määratud ühele retoorilisele kategooriale, ning sisaldab nelja kriteeriumiga truudusfiltrit ja tõendipõhist moderaatorit, kes lisab voorude vahel vastaspoole korpusest pärinevaid vastutõendeid.

Töövoogu hinnati SCALE märgendamise usaldusväärsuse, debativastuste allikatruuduse ning 91 kõnelejate paari lõikes positsioonide ja vaadeldud konvergentsi vahelise seose alusel.

Võtmesõnad: maskuliinsuse diskursus, suured keelemudelid, otsingutel põhinev genereerimine, mitme agendi debatt, SCALE märgendamine, arvutuslingvistika, YouTube

CERCS: P176 Tehisintellekt

Table of Contents

1. Introduction

Public discourse around masculinity has become increasingly polarised over the past decade, with YouTube and adjacent platforms hosting a wide range of thought leaders — from manosphere-aligned commentary on dominance and gender hierarchy to therapeutic and progressive accounts of masculine vulnerability. The discourse reproduces readily across linguistic and cultural contexts, including in small media markets such as Estonia [1]. Systematic analysis of how individual speakers within this ecosystem frame masculinity is therefore of scholarly as well as societal interest.

Computational approaches to discourse of this kind — for instance word-embedding analyses — capture aggregate patterns but flatten the argumentative structure through which masculinity ideals are defended and negotiated. Persona-conditioned large language models offer a more dialogically rich alternative but tend to drift from the speaker they are meant to represent, amplifying stereotypical attributes when conditioned on group identity and explaining only a small fraction of variance in subjective tasks.

This thesis addresses that gap by combining three techniques developed largely independently: retrieval-augmented generation (RAG) [2] grounds generated text in retrieved source passages; multi-agent debate [3] elicits contrasts between positions by forcing model instances to argue against each other; and the SCALE framework [4] simulates the collaborative content-analysis workflow used in qualitative social science. The pipeline developed here integrates all three: SCALE annotates a corpus of YouTube transcripts for rhetorical function, those annotations guide retrieval inside a RAG component, and the retrieved passages condition a moderated multi-agent debate.

The objective of the thesis is to design, implement, and validate this pipeline. The validation has two parts: a *reliability* question — whether the SCALE annotation produces consistent rhetorical codes across runs and agents — and a *validity* question — whether the debate system generates responses faithful to each speaker's documented positions rather than to the model's general assumptions about them. Only when both are answered affirmatively can the system's outputs support substantive claims about the discourse itself. To this end, the pipeline is run over a corpus of fourteen English-language YouTube thought leaders, with five debate questions spanning identity, partnership, critique, development, and aspiration in contemporary masculinity discourse. The computer-science contribution of the thesis is the pipeline itself; the masculinity discourse serves as both motivation and testbed.

The thesis is organised as follows. Chapter 2 reviews the theoretical background on masculinity ideals, large language models as persona modellers, and the SCALE annotation framework. Chapter 3 describes the data: speaker selection, transcript extraction, corpus statistics, and the debate questions. Chapter 4 documents the methodology behind the text-processing and vector-storage layer, the SCALE annotation pipeline and its HPC deployment, and the debate system with its retrieval, prompting, filtering, and moderation components. Chapter 5 reports the reliability of the multi-agent annotation, the validity of the debate system, and the substantive findings on masculinity discourse. Limitations chapter discusses the constraints that qualify the findings and chapter 7 concludes. Appendix A lists the source videos that make up the corpus.

Throughout the development of this thesis, Claude (Anthropic) was used as an interactive assistant across several distinct tasks: code scaffolding and debugging during pipeline implementation, structural and editorial feedback on chapter drafts, exploratory literature search and source summarisation, and prose refinement. All theoretical claims, methodological decisions, citation selection, and final wording were made by the author, and any AI-suggested text that entered the thesis was verified against the underlying sources before inclusion. Zotero with an integrated reference-formatting model was used for citation management.

2. Theoretical Background

2.1 Research on Masculinity Ideals

Earlier essentialist accounts of masculinity have given way to constructivist frameworks that treat it as a set of socially constructed expectations surrounding male roles [5]. In its most influential formulation, Connell [6] defines *hegemonic masculinity* as the culturally dominant configuration of male practice that legitimises men's authority while subordinating alternative masculinities. The concept does not describe a fixed character type but a shifting ideal: in contemporary Western culture it often manifests as the expectation that men should be economically self-sufficient, emotionally stoic, physically strong, and heterosexually dominant [6].

A complementary perspective comes from social role theory, which argues that gendered expectations originate in the historically different social positions occupied by men and women: men have predominantly occupied breadwinner and public-sphere roles and are culturally associated with *agentic* qualities such as independence, assertiveness, and achievement orientation, while women's concentration in caregiving roles has produced associations with *communal* qualities such as warmth and emotional expressiveness [7]. Bhatia and Bhatia [8] tracked these associations across a century of American English text using word embeddings and found a notable asymmetry: associations between women and communal traits have weakened considerably, reflecting women's increased participation in professional and public life, while associations between men and agentic traits have remained largely unchanged. Masculine ideals appear culturally more resistant to revision than feminine ones, which helps explain the intensity of contemporary debates over what masculinity should look like.

In the digital era, social media platforms have become primary sites where masculinity ideals are constructed, commercialised, and consumed. A loose ecosystem of online communities organised around men's grievances, antifeminism, and, in more extreme forms, explicit misogyny has come to be referred to as the *manosphere* [1], [9], and the figures who articulate masculinity ideals to large online audiences are commonly termed *manfluencers* [1]. Within the YouTube ecosystem specifically, Kompatsiaris [9] analyses NoFap lifestyle gurus and shows how the platform produces what he terms "spiritpreneurial masculinity": a hybrid ideal that blends entrepreneurial self-optimisation with spiritual practice — even within a single

ideological cluster the platform generates novel masculinity configurations that do not map neatly onto traditional categories. Gram et al. [10] analysed testosterone-related content on Instagram and TikTok and found that influencers frame low testosterone as a crisis of manhood, constructing a binary opposition between being a "real man" and being feminine. Lott, Murumaa-Mengel, and Marling [1] document how this phenomenon operates even in small media markets: Estonian TikTok manfluencers adopt internationally circulating manosphere ideologies with minimal local adaptation. Together these studies illustrate how contemporary platform-based masculinity discourse converges on a recurring set of contested questions — what makes a "real man," what is wrong with men today, what boys should be taught to aspire to, and whom they should look up to — questions that form the substantive territory examined computationally in this thesis through the five research questions specified in Section 3.4.

Studying these discourses computationally requires analytical precision. Lucy [11] warns that researchers frequently commit "conceptual slippage" when applying Connell's framework to online masculinity communities: dominant or dominating masculine performances — such as aggression, sexual prowess, or physical strength — are labelled hegemonic without demonstrating how these traits actually legitimise unequal gender relations [11]. The implication for the present work is methodological: any analysis of masculinity discourse must attend to whether specific discursive positions function to legitimise masculine hierarchies or merely to perform dominant masculine behaviour. This distinction motivates the codebook's instruction to classify by rhetorical function rather than tone (Section 4.3.1) and frames the interpretation of the per-speaker rhetorical fingerprints reported in Section 5.1.4.

Computational approaches such as word-embedding analysis [8] capture aggregate patterns at scale but flatten the argumentative structure through which masculinity ideals are defended and negotiated; large language models open up a more dialogically rich alternative [12]. The choice to use LLMs as persona-conditioned debaters in this thesis is motivated by the observation that masculinity ideals are not collections of associated words but argumentative positions — defended, attacked, nuanced, and negotiated in dialogue. The remainder of this chapter develops the technical apparatus required to make such a debate-based analysis tractable: persona simulation grounded in retrieved evidence (Section 2.2) and a multi-agent annotation framework that codes the retrieved material for rhetorical function (Section 2.3).

2.2 Large Language Models as Persona Modellers

2.2.1 The Simulated Persona

A central premise of this thesis is that a large language model can be conditioned to approximate the discursive behaviour of a specific real-world individual by generating responses grounded in their recorded speech. Shanahan, McDonell, and Reynolds [13] provide the theoretical basis for this approach, arguing that LLMs are best understood not as entities with fixed beliefs but as non-deterministic simulators.

Persona prompting alone, however, is a comparatively weak conditioning mechanism. Hu and Collier [14] found that persona variables explain less than 10% of the variance in most subjective NLP tasks, suggesting that telling a model who it is does relatively little to determine what it says. A separate and more troubling failure mode is documented by Cheng et al. [15], who show that LLMs systematically amplify stereotypical attributes of social groups when prompted to simulate their members. The model produces what it has learned to *expect* from a category rather than what any specific member of that category has actually said.

Both findings carry direct implications for the present system. Persona prompting by itself — telling the model "you are Andrew Tate" — is insufficient to produce outputs that faithfully represent the speaker's actual views; the model will instead produce what it "thinks" Tate would say, drawing on its training data in ways that may exaggerate, flatten, or distort the speaker's position. This motivates the two grounding mechanisms central to the methodology: retrieval-augmented generation (Section 2.2.2) anchors each generated response in specific transcript chunks from the speaker's own corpus, and a faithfulness filter (Section 4.4.4) rejects responses whose claims cannot be traced back to those chunks. Together they shift the conditioning load from the persona prompt to the source material itself.

2.2.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation augments a language model's generation capabilities with an external retrieval component, allowing it to condition outputs on specific, verifiable source documents rather than relying solely on parametric knowledge. Lewis et al. [2] introduced the RAG framework, in which a retriever identifies the most relevant passages from a non-parametric document index and the generator produces output conditioned on both the query and the retrieved passages. The framework achieved state-of-the-art results on knowledge-

intensive tasks and established retrieval grounding as a substantial improvement over parametric-only generation [2].

The RAG paradigm has since become a dominant approach in applied NLP. Gao et al. [16] survey the field and identify three evolutionary stages: *Naive RAG*, a simple retrieve-then-read pipeline; *Advanced RAG*, which adds pre- and post-retrieval optimisations such as query rewriting, re-ranking, and hybrid search; and *Modular RAG*, which decomposes the pipeline into interchangeable components. The system developed here falls between Advanced and Modular RAG: it uses a hybrid retrieval strategy combining a semantically annotated database with a full unannotated corpus, and applies a post-generation faithfulness filter that can trigger re-retrieval when the initial retrieval is judged insufficient.

The connection between RAG and persona-conditioned dialogue is well established. Shuster et al. [17] demonstrated that augmenting dialogue models with retrieval over a knowledge source substantially reduces hallucination in conversation. Huang et al. [18] address the persona-specific case directly: their LAPDOG framework retrieves supplementary context to enrich short persona profiles — typically four to five sentences in benchmark datasets such as CONVAI2 — and shows that retrieved content can compensate for the thin descriptive signal of a persona prompt alone. Both findings apply directly to the present system, in which each debate bot retrieves exclusively from its own speaker's transcript corpus, using the transcripts themselves as the supplementary source that anchors the otherwise minimal persona prompt.

Two practical design decisions in any RAG system are the vector database and the chunking strategy. The present system stores chunks in ChromaDB [19], an open-source embedding database that supports metadata filtering — a feature exploited by the retrieval system to filter and boost candidates by their rhetorical category (Section 4.4.2). ChromaDB was selected for its native metadata filtering support and lower configuration overhead, which are more relevant than raw throughput for the corpus sizes involved here. For chunking, Qu et al. [20] found that the computational cost of semantic chunking — segmenting at topical boundaries detected by embedding similarity — is not consistently justified by performance gains over simple fixed-size chunking; the present system therefore uses fixed-size chunking with overlap (Section 4.2).

2.2.3 Multi-Agent Debate and the Moderator Role

Multi-agent debate (MAD) is a framework in which multiple LLM instances engage in structured argumentation, each agent defending a position and responding to the others' arguments across several rounds. Du et al. [3] provided the first large-scale empirical evidence that multi-agent debate improves LLM factuality and reasoning: in their framework, multiple model instances independently generate initial responses and then iteratively refine their answers over several rounds by reading and responding to other agents' outputs.

Liang et al. [21] formalised the Multi-Agent Debate framework and introduced a dedicated *judge* or *moderator* role. In their architecture, debating agents present opposing viewpoints while a separate judge agent evaluates the arguments, provides guidance, and determines a final verdict. Without external guidance, agents may converge toward bland consensus or become stuck in repetitive disagreement. The moderator in the present system serves three functions: it analyses each round's arguments and provides structured feedback that guides the next round, it identifies claims that lack evidential support and triggers evidence re-retrieval, and it produces a final synthesis comparing the debaters' positions (Section 4.4.5).

Whether multi-agent debate reliably improves on simpler baselines remains contested. Smit et al. [22] conducted a systematic comparison of MAD strategies and found that debate does not consistently outperform simpler approaches such as self-consistency; their analysis suggests that debate is most beneficial for tasks requiring *diverse perspectives* rather than convergence toward a single correct answer. This characterisation matches the present application closely, where the analytical goal is to surface ideological contrasts between speakers rather than identify a correct answer to questions like "what makes a good husband."

Wu et al. [23] document that *majority pressure* — the suppression of minority positions during multi-agent debate — can lead to false consensus rather than genuine deliberative reasoning. This dynamic is directly analogous to the *positional drift* problem observed during the development of the present system, in which a bot representing one speaker gradually adopted the opposing speaker's position under moderator prompting. The present system mitigates this risk through two mechanisms: retrieval grounding restricts each bot to its own speaker's corpus, and the faithfulness filter rejects responses that break character with respect to the retrieved source material (Section 4.4.4).

2.3 Annotation with the SCALE Framework

Content analysis is the systematic coding of textual data into predefined categories that traditionally relies on multiple human coders, with inter-annotator agreement serving as the primary measure of coding reliability. It is labour-intensive and does not scale easily to the large corpora typical of computational studies [12], [24]. Recent work has demonstrated that large language models can perform annotation tasks at a level comparable to or exceeding that of trained crowd workers. Gilardi, Alizadeh, and Kubli [24] found that ChatGPT outperformed MTurk crowd workers on several annotation tasks including relevance classification, stance detection, and topic coding at approximately twenty times lower cost.

Most LLM annotation approaches, however, treat the model as a single annotator applying a fixed codebook, missing a defining feature of social-science content analysis: the collaborative refinement of the codebook through discussion of ambiguous cases. The SCALE framework (Scalable Collaborative Annotation with LLM Experts), introduced by Zhao et al. [4], simulates this full workflow using multiple LLM agents. In the first stage, agents independently code a set of textual items using an initial codebook. In the second, they engage in a structured discussion about disagreements, presenting their reasoning and attempting to reach consensus — mirroring the adjudication discussions that human coding teams conduct. In the third, the codebook itself is revised based on insights from the discussion, and the process iterates. Critically, SCALE includes an explicit mechanism for *human intervention*: a human researcher can review the agents' discussion, override decisions, and modify the codebook at any point, ensuring that domain expertise guides the process [4].

The choice to use a multi-agent annotation framework rather than a single LLM annotator is motivated by the same logic that underlies the debate system itself: discourse of this kind is inherently ambiguous, and reasonable annotators may disagree about how to code a given passage. SCALE's collaborative discussion mechanism surfaces these ambiguities rather than concealing them behind a single annotator's choice. The codebook adapted for the present thesis — classifying transcript chunks into three rhetorical categories that support the debate system's category-aware retrieval — and the autonomous-mode deviation from Zhao et al.'s human-intervention design are documented in Section 4.3.

3. Data

3.1 Speaker Selection

The study analyses the masculinity discourse of fourteen English-language YouTube thought leaders. Speakers were selected through purposive sampling to maximise ideological contrast across the sample while ensuring that each had a sufficiently large and publicly accessible body of long-form spoken material on the topic.

Selection followed three criteria: (1) at least ten long-form videos (60+ minutes, or 30+ minutes for videos closely related to the research topic where longer ones were unavailable) in which masculinity-related topics are discussed substantively; (2) a clearly identifiable position within the masculinity discourse landscape, such that the speaker's inclusion creates productive ideological contrast with the rest of the sample; and (3) transcripts of sufficient quality obtainable through YouTube's captioning system.

An initial long list of eleven candidates was generated with ChatGPT (GPT-4), used as an exploratory research tool rather than as an analytical instrument [12]. Three further speakers were added manually based on podcast cross-referencing: when a candidate had been interviewed by the same hosts as speakers already in the sample, this indicated discursive adjacency and potential for meaningful comparison. The manual additions also updated the sample's temporal coverage, since several GPT-generated candidates had been more active in public discourse some years before data collection. The final sample of fourteen speakers represents four broad discursive orientations within the contemporary masculinity debate. The complete list of source videos is provided in Appendix A.

The first orientation is radical masculinity and "Red Pill" ideology, represented by Andrew Tate and the Fresh & Fit podcast. Both speakers articulate what Lucy [11] characterises as dominant or dominating masculine performance emphasising status hierarchy, sexual conquest, and emotional suppression and operate within the broader manosphere ecosystem documented by Kompatsiaris [9] and Lott et al. [1]. Tate promotes an aggressive, status-driven model of alpha masculinity, while Fresh & Fit applies evolutionary-psychology framing to dating-market dynamics. Their discourse exemplifies the agentic-trait monopoly that Bhatia and Bhatia [8] identified as culturally resistant to change, and reproduces the "real man versus feminine" binary that Gram et al. [10] documented in adjacent influencer ecosystems.

The second orientation is discipline, stoicism, and self-optimisation, represented by Jocko Willink, Andrew Huberman, Ryan Holiday, and Naval Ravikant. These speakers reframe traditionally agentic traits like self-sufficiency, control and mastery through registers of virtue ethics, neuroscience, or philosophical productivity rather than dominance. Willink articulates masculinity through military discipline and stoic endurance, Huberman through neuroscientific self-optimisation, Holiday through ancient Stoic philosophy, and Ravikant through intellectual sovereignty and philosophical self-mastery. The orientation is proximate to the hybrid configuration Kompatsiaris [9] terms spiritpreneurial masculinity, in which entrepreneurial self-cultivation is fused with disciplined or contemplative practice. Compared to the radical orientation, this group retains the agentic core of the traditional masculine configuration described by Connell [6] but routes it through projects of internal cultivation rather than external dominance.

The third orientation is psychology, relationships, and mental health, represented by Jordan Peterson, Esther Perel, John Gottman, and Dr. K (HealthyGamerGG). The group is internally heterogeneous: Peterson frames masculinity through responsibility, competence hierarchies, and Jungian archetypes — defending what is partly a traditional configuration in psychological language — while Perel centres emotional intelligence and relational dynamics, and Gottman contributes empirically grounded relationship psychology. Dr. K bridges Western clinical psychiatry with Eastern contemplative traditions in addressing young men's mental health. What unites the group is the treatment of masculinity primarily as a clinical or developmental matter, foregrounding communal and relational competencies.

The fourth orientation is social criticism and alternative perspectives, represented by Brené Brown, ContraPoints (Natalie Wynn), Destiny (Steven Bonnell), and Alison Armstrong. These speakers contest the traditional masculine configuration from different angles. Brown reframes vulnerability as a form of strength, directly challenging the emotional-stoicism norm; ContraPoints offers a feminist video-essay critique of masculinity as social performance; Destiny applies internal-rationalist deconstruction to manosphere claims; and Armstrong frames gendered difference as relational and developmental rather than hierarchical. As a group, they engage explicitly with the constructivist understanding of masculinity outlined by Connell and Messerschmidt [5].

These four orientations were established a priori on the basis of each speaker's public positioning and serve as a working hypothesis for the structure of the sample rather than as an

empirical claim. Whether the rhetorical patterns produced by the annotation pipeline support this grouping is examined in Section 5.3.

3.2 Source Material and Transcript Extraction

All source material is publicly available YouTube content from public figures who publish their discourse intentionally for public consumption. For each speaker, approximately ten YouTube videos were selected as the source corpus. Source material was restricted to interviews and monologues so formats in which the studied speaker holds the substantive speaking role and develops extended arguments. Multi-speaker debate-format videos would in principle have been the most natural fit for a project on debate modelling, but were excluded because YouTube's automatic captioning system performs no speaker diarisation: utterances from different speakers would be conflated in the transcript, contaminating one speaker's corpus with another's. For podcast-format speakers such as Fresh & Fit and Destiny, videos were selected in which the target speaker holds the primary speaking role.

Transcripts were obtained using YouTube's captioning system, accessed programmatically through the youtube-transcript-api Python library [25]. Of the 140 transcripts in the corpus, 129 use YouTube's automatically generated captions, while 11 have manually uploaded caption tracks provided by the video creators. No manual correction of ASR errors was performed, as the downstream pipeline operates on semantic similarity at the passage level rather than exact lexical matching. Three of the 140 transcripts were inadvertently scraped twice during corpus construction; the database therefore contains 140 transcripts derived from 137 unique videos. The chunking, embedding, and storage of these transcripts are described in Section 4.2.

3.3 Corpus Statistics

Table 1 summarises the corpus statistics for each speaker. The retrieval database comprises 23,776 chunks built from 140 transcripts representing 137 unique videos and approximately 3.02 million words across fourteen speakers. Estimated chunk counts are computed using the chunking parameters described in Section 4.2 and are reported here as an indicator of corpus size in computational terms. The substantial variation in corpus size between speakers reflects primarily differences in speaking rate, interview length, and verbosity, with smaller contributions from variation in the number of source videos. This nearly threefold range in corpus size means that speakers with larger corpora offer greater thematic coverage per query, while speakers with smaller corpora may exhibit higher retrieval overlap across queries. The

implications of this imbalance for annotation and debate generation are addressed in the Limitations chapter.

Table 1. Corpus statistics per speaker. The Unique Videos column reports unique source videos. Word and chunk counts reflect the corpus as constructed (140 transcripts from 137 unique videos; three videos are represented twice in the database — see Section 3.2). Estimated chunk counts assume the chunking parameters described in Section 4.2. Transcript type indicates how many of each speaker's videos have auto-generated versus manually uploaded captions. Speakers are grouped by discursive orientation.

Speaker	Unique Videos	Words	Chunks	Transcripts	Orientation
Andrew Tate	10	280,792	2,250	10 auto	Radical
Fresh & Fit	9	280,499	2,023	10 auto	Radical
Jocko Willink	10	311,285	2,493	9 auto, 1 manual	Discipline
Andrew Huberman	10	327,043	2,620	9 auto, 1 manual	Discipline
Ryan Holiday	10	193,503	1,551	10 auto	Discipline
Naval Ravikant	9	251,697	1,821	10 auto	Discipline
Jordan Peterson	8	160,032	1,237	9 auto	Psychology
Esther Perel	11	175,090	1,406	8 auto, 3 manual	Psychology
John Gottman	10	117,722	944	9 auto, 1 manual	Psychology
Dr. K (HealthyGamerGG)	10	255,267	2,044	10 auto	Psychology
Brené Brown	10	132,358	1,062	9 auto, 1 manual	Social criticism
ContraPoints	10	114,530	920	6 auto, 4 manual	Social criticism
Destiny	10	268,728	2,152	10 auto	Social criticism
Alison Armstrong	10	156,141	1,253	10 auto	Social criticism
Total	**137**	**3,024,687**	**23,776**	**129 auto, 11 manual**	—

3.4 Debate Questions

Five debate questions guide the analysis of each speaker's corpus:

1. What does it mean to be a real man?
2. What qualities make a good husband?
3. What are the worst qualities of most men today?
4. What qualities should be encouraged in boys growing up?
5. Who should young men look up to?

These questions were deliberately formulated in a direct, colloquial register rather than in formal academic language. Because they are used as semantic search queries against a corpus of spoken discourse (Section 4.4.2), phrasing them in the same register as the source material improves retrieval precision by maximising embedding similarity between the query and relevant transcript passages.

The five questions span the core thematic territory of contemporary masculinity discourse — identity (RQ1), partnership (RQ2), critique (RQ3), development (RQ4), and aspiration (RQ5) — while being specific enough to elicit substantively different responses from speakers with different ideological commitments. They are designed to surface contrasting positions across the speaker sample: a question such as "What qualities make a good husband?" invites different answers from a relationship therapist (Gottman, Perel), a stoic-discipline advocate (Willink, Holiday), and a Red Pill commentator (Tate, Fresh & Fit), and the analytical purpose of the system is to make these contrasts visible.

These same five questions serve two distinct functions in the pipeline. They direct the SCALE annotation process by determining which transcript passages are annotated for rhetorical function (Section 4.3), and they define the topics on which the debate system is run (Section 4.4).

4. Methodology

This chapter describes the system architecture, its components, and the design decisions that shaped them. Figure 1 provides a high-level overview of the pipeline, from raw transcripts through the SCALE annotation framework to the final debate output. Section 4.1 presents the analytical purpose that guides every architectural decision. Section 4.2 describes how raw transcripts are converted into a searchable vector representation. Section 4.3 describes the SCALE multi-agent annotation framework, including the codebook, agent personas, workflow, and the computational infrastructure used for the production annotation run. Section 4.4 describes the debate system itself, including round structure, retrieval logic, prompt construction, the faithfulness filter, and the moderation layer. The complete implementation is publicly available at https://github.com/KimLiliiTamm/masculinity-debate.

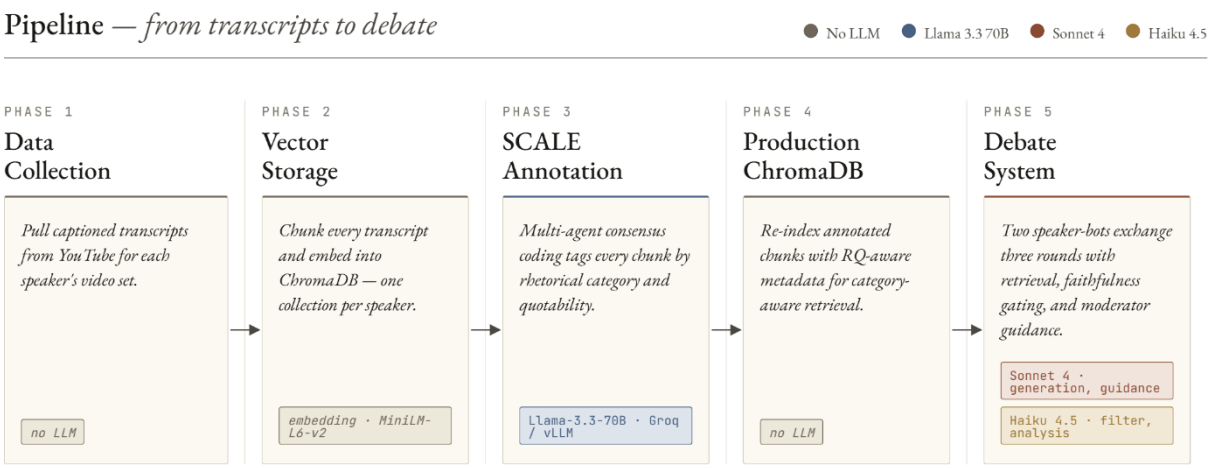

Figure 1. High-level overview of the pipeline, from raw transcripts through SCALE annotation to the moderated debate output.

4.1 General Architecture

The pipeline is structured around a single design requirement: every claim attributed to a speaker in the generated debate must be traceable to a specific segment of that speaker's transcribed discourse. The chunking strategy, the annotation pipeline, the retrieval logic, and the faithfulness filter are each motivated by this requirement.

Three distinct language models serve different roles in the pipeline. Llama 3.3 70B Instruct [26] handles the SCALE annotation phase, where the priority is cost-efficient processing of 7,500 transcript chunks (15 speaker-slots × 5 research questions × 100 chunks; one slot was a duplicate, leaving 7,000 unique chunks for downstream analysis — see Section 4.3.6) on the

university's HPC cluster. Claude Sonnet [27] generates debate responses and moderator guidance, where output quality and rhetorical sophistication are critical. Claude Haiku [28] evaluates response quality through the faithfulness filter and performs round analysis, where speed and cost efficiency matter more than generative capability. This separation of models by function reflects a deliberate cost–quality trade-off: annotation and evaluation tasks do not require the same model capacity as open-ended generation.

4.2 Text Processing and Vector Storage

The plain-text transcripts produced from the YouTube captions (Section 3.2) are segmented into fixed-size chunks of approximately 150 words each, with an overlap of 25 words between consecutive chunks. The overlap reduces the risk of splitting a coherent claim across two fragments. The 150-word size was chosen empirically based on preliminary experiments with the debate pipeline: shorter chunks (e.g., 50–100 words) frequently lacked sufficient argumentative content, while longer chunks (e.g., 300+ words) often mixed multiple topics, degrading retrieval precision. The decision to use fixed-size rather than semantic chunking follows the finding by Qu et al. [20] discussed in Section 2.2.2.

Each chunk is embedded using the all-MiniLM-L6-v2 sentence-transformer model [29], a lightweight transformer that maps text to 384-dimensional dense vectors. This model was selected for two reasons. Its 256-token maximum input length is suitable for 150-word chunks (approximately 200 tokens), and it can be run locally without GPU acceleration.

Embedded chunks are stored in ChromaDB [19] using one collection per speaker, with each chunk tagged with metadata identifying the speaker, source video, and chunk index. The full database contains all chunks from all videos without rhetorical annotation, ranging from approximately 920 to 2,620 chunks per speaker. The strict separation of speakers into individual collections enforces an information boundary central to the system's design: each debate bot can only retrieve material from its own speaker's collection, preventing cross-contamination of personas and complementing the faithfulness filter (Section 4.4.4).

The system maintains two ChromaDB databases. The full database described above provides broad coverage. A production database (Section 4.3.5) contains the subset of chunks annotated through the SCALE pipeline and carries rhetorical category metadata; retrieval queries it first for rhetorically classified material and falls back to the full database when the annotated subset does not provide sufficient coverage. This hybrid architecture addresses a tension between cost

and coverage: SCALE annotation is expensive enough that applying it to the full 23,776-chunk corpus would be prohibitive, but restricting retrieval exclusively to the annotated subset produces chunk starvation for speakers whose discourse does not naturally distribute across all rhetorical categories.

4.3 SCALE Annotation

Each transcript chunk must be tagged with its rhetorical function before the debate's retrieval strategy can use it. The system adapts the SCALE framework described in Section 2.3 [4] for this purpose. Two features of the present implementation are worth flagging at the outset. First, the codebook (Section 4.3.1) is specific to masculinity discourse rather than a general-purpose coding scheme. Second, annotation runs separately for each of the five research questions: rhetorical function is treated as relative to the analytical question being asked, so the same chunk may receive different codes when coded against different questions.

4.3.1 Codebook Design

The codebook defines three primary rhetorical categories — *Core Dogma*, *Provocation*, and *Rhetorical Evidence* — and assigns three annotations to every chunk: a primary code, a secondary code, and a quotability rating. Core Dogma identifies statements that assert fundamental beliefs about masculinity, society, or human nature. Provocation identifies statements that challenge opposing viewpoints or use inflammatory language to assert dominance. Rhetorical Evidence identifies statements that support a position through anecdotes, clinical examples, biological claims, or metaphors. The codebook instructs agents to classify by rhetorical *function* rather than tone, since a calmly stated belief and an aggressively stated belief can share the same function despite differing in delivery — a distinction motivated by the conceptual-slippage problem discussed in Section 2.1. The full codebook text appears in Appendix B.1.

The secondary code captures a subordinate rhetorical function using the same three categories, or 0 if the chunk serves only one. This dual-coding scheme proved important for retrieval: during development, restricting retrieval to primary codes alone caused chunk starvation for speakers whose discourse concentrates in one or two categories. The secondary code provides an additional retrieval pathway, exploited by the boost mechanism described in Section 4.4.2.

The quotability dimension marks whether a chunk contains a memorable, self-contained phrase that would function effectively as a direct quote in a debate context. Chunks rated Q receive a retrieval distance bonus during debate generation, increasing the likelihood that the bot's response incorporates the speaker's most distinctive formulations.

4.3.2 Agent Personas

The SCALE implementation uses three LLM agents acting as annotators, each assigned an academic persona aligned with one of the three codebook categories. Dr. Sterling is a political sociologist primed to detect Core Dogma through his focus on ideology and hierarchy; Dr. Hayes is a gender studies scholar with a clinical psychology background, attuned to claims about gender and emotional norms; Dr. Vance is a rhetoric and persuasion specialist oriented toward Provocation and argumentative strategy. Full persona system prompts appear in Appendix B.2.

This alignment between personas and codebook categories is a deliberate design choice that increases the probability of productive disagreement: when agents disagree, they tend to disagree for substantive disciplinary reasons rather than random variation. The alignment also introduces a bias risk — each agent is primed toward a particular category — which is mitigated by the discussion mechanism and majority voting.

4.3.3 Annotation Workflow

The annotation process follows a two-phase structure for each batch of text chunks: independent coding followed by consensus-seeking discussion. In the coding phase, all three agents independently classify each chunk; they operate in parallel and do not see each other's responses. Each agent provides a reasoned explanation quoting the sentence that justifies its classification and states its final numerical codes at the end of the response. A separate judge agent then compares the three annotations. If they match, the coding-phase responses are accepted as final; if any disagreement is detected, the chunk enters the discussion phase, in which each agent receives the other agents' responses and may revise its own. Discussion continues for up to two rounds, ending early if consensus is reached (Figure 2). The two-round cap prevents indefinite cycling on genuinely ambiguous chunks — a pattern observed during development. The exact coding, discussion, and judge prompts appear in Appendix B.3.

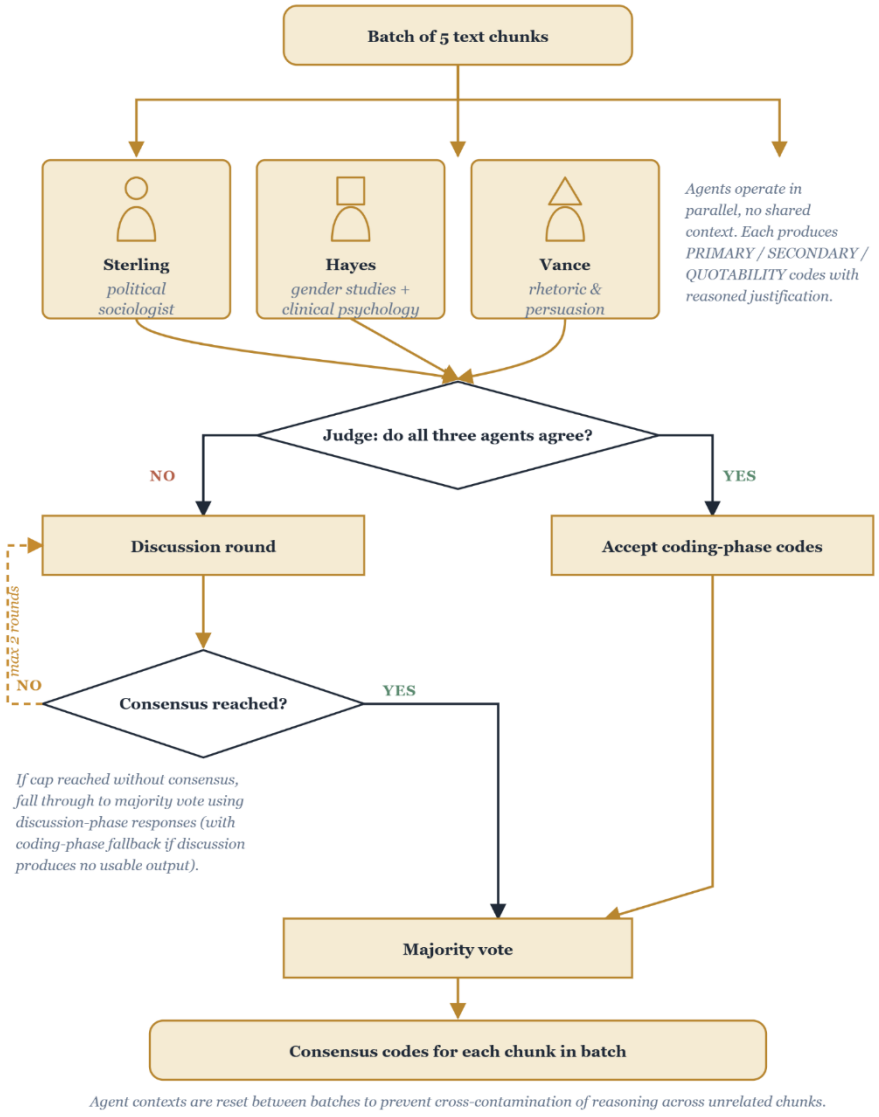

Figure 2. The SCALE annotation workflow for one batch of chunks.

The implementation operates in fully autonomous mode: the human intervention mechanism described in the original SCALE framework [4] is not used. No human researcher reviews agent discussions, overrides coding decisions, or modifies the codebook during annotation. This decision was made to enable annotation of the corpus (7,000 chunks across 14 speakers and 5 research questions) within the time and resource constraints of a bachelor's thesis. The trade-off is that annotation quality depends entirely on the agents' ability to apply the codebook correctly without correction; the methodology accounts for this by using majority voting and by accepting that some boundary cases will receive inconsistent codes, an outcome mirroring genuine ambiguity in the underlying discourse. The consequences for reliability are reported in Section 5.1.

4.3.4 Code Extraction and Consensus

The raw output of each SCALE run is a JSON log containing each agent's full reasoning and final codes for every chunk. A post-processing script extracts the numerical codes from these responses using pattern matching, since agents do not always format their answers identically. The extraction logic attempts six regex patterns in descending order of specificity, from structured formats such as "PRIMARY: 1 SECONDARY: 2 QUOTABILITY: Q" to inline formats such as "Final numerical answer: 1, 2, Q." This flexibility was necessary because the Llama 3.3 model used for annotation does not reliably adhere to a single output format across thousands of responses, despite explicit formatting instructions in the prompt.

For each chunk, the primary code, secondary code, and quotability rating are determined by majority vote across the three agents. If the judge determined that agents agreed during the coding phase, the coding phase responses are used; if disagreement triggered discussion, the final discussion-phase responses are used instead, with a fallback to coding-phase responses if the discussion produced no usable output.

4.3.5 Production Database Construction

The consensus codes produced by the SCALE pipeline are stored in a structured JSON file keyed by chunk identifier and research question, which is then used to build the production ChromaDB. Because each chunk may be annotated against multiple research questions and may receive different rhetorical codes depending on the question context, annotations are stored in a per-research-question namespace rather than as a single global code. A representative chunk record for research question 1, for example, would carry the fields rq1_primary=1, rq1_category="Core Dogma", rq1_secondary=3, and rq1_quotability="Q". After deduplication across research questions, the production database contains approximately 250 to 340 unique chunks per speaker.

4.3.6 Computational Infrastructure

For cost efficiency, the SCALE annotation pipeline was deployed on the University of Tartu Rocket HPC cluster [30], which made it possible to run the Llama 3.3 70B model locally at no per-token cost. Earlier pipeline validation used the Groq cloud API hosting the same model.

Model serving. The HPC deployment runs Llama 3.3 70B Instruct locally using vLLM [31], distributed across two NVIDIA H200 GPUs via tensor parallelism, with a context window of 32,768 tokens. The context window size was determined empirically by replaying message accumulation patterns from earlier Groq API logs and confirming that 32,768 tokens accommodated the peak coding-and-discussion-phase usage.

Production-run optimisations. Two configuration changes were applied based on validation-phase findings. Codebook evolution — a feature of the original SCALE framework in which agents collaboratively revise the codebook between batches — was disabled for production after being used to develop the codebook itself in a preliminary pilot run over 30 chunks from each of five speakers (Andrew Tate, Jocko Willink, Jordan Peterson, Esther Perel, and Andrew Huberman). The codebook that emerged from the pilot was frozen as the version reproduced in Appendix B.1 and applied uniformly across all 14 speakers in the production run, saving tokens that would otherwise be spent on between-batch revisions. Discussion length was capped at 200 words per agent per round, reducing discussion-phase token consumption by approximately 60–70 percent without measurably affecting inter-agent agreement. The annotation script processed 15 speaker-slots in parallel; one slot was inadvertently populated with a second copy of another speaker's transcripts, producing 7,500 annotated chunks of which 7,000 correspond to 14 unique speakers. The duplicate slot's annotations are used as a test-retest sample in Section 5.1.2 and excluded from all downstream analysis. The final annotation of all 14 speakers ran for approximately 15 hours on a two-GPU node, costing an estimated 19 euros at the university's internal HPC rates.

4.4 Debate System

The debate system generates structured two-speaker debates grounded in real transcript material through retrieval-augmented generation. Each debate pairs two speakers from the corpus, poses one of the five research questions introduced in Section 3.4, and runs for three rounds. The system enforces a strict constraint: every substantive claim in a bot's response must be traceable to a specific chunk from that speaker's transcript corpus. All generated debate outputs are treated and presented as LLM simulations conditioned on transcript evidence, not as direct quotes or genuine opinions of the speakers. This section describes the six components that implement this constraint — the round strategy, the retrieval mechanism, the prompt construction, the quality filter, the moderation layer, and the orchestration logic that connects them (Figure 3).

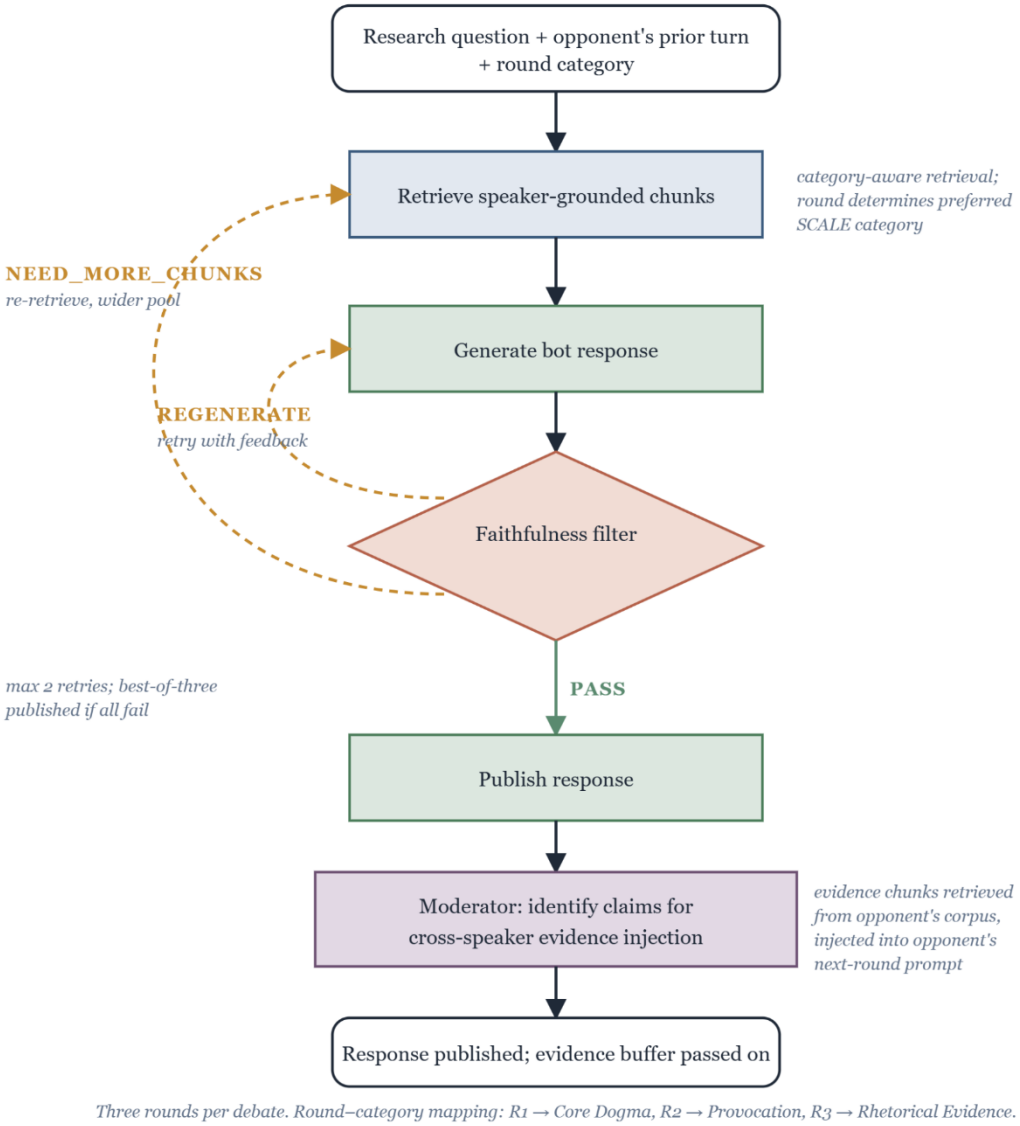

Figure 3. Per-turn debate generation flow, including the faithfulness-filter routing.

4.4.1 Round Structure and Rhetorical Strategy

Each debate consists of three rounds, and each round is mapped to one of the three SCALE annotation categories described in Section 4.3.1. Round 1 is assigned to *Core Dogma*: both bots lead with their speaker's fundamental beliefs and worldview, asserting philosophical positions about masculinity, society, and human nature. Round 2 is assigned to *Provocation*: both bots challenge the opponent's framing directly, attacking weak points, questioning assumptions, and asserting dominance over the narrative. Round 3 is assigned to *Rhetorical*

Evidence: both bots ground their arguments in concrete examples — personal stories, data claims, biological arguments, metaphors, or appeals to authority.

This mapping serves two purposes. First, it structures the debate as a progression from ideological positioning through confrontation to evidential justification, mirroring a natural argumentative arc. Second, the round's assigned category is used as a metadata filter in the retrieval step (Section 4.4.2), so each round draws on the appropriate rhetorical material from the SCALE-annotated production database.

Within each round, Bot 1 always responds first, followed by Bot 2. This fixed ordering means that Bot 2 always has access to Bot 1's response in the current round, while Bot 1 does not see Bot 2's response until the next round. In Round 1, both bots respond to the research question directly. In Rounds 2 and 3, each bot receives the opponent's most recent response and is instructed to dedicate approximately 65 percent of its response to addressing specific claims the opponent made, with the remaining 35 percent devoted to advancing its own position using new source material. This ratio was calibrated during development to prevent two failure modes: pure monologue (where bots ignore each other entirely) and pure reaction (where bots abandon their speaker's voice to rebut the opponent point by point).

4.4.2 Category-Aware Retrieval

The retrieval system queries the dual-database architecture described in Section 4.2, combining results from the SCALE-annotated production database and the full transcript corpus. For each turn, retrieval proceeds in six steps:

1. **Formulate the search query.** The query is constructed differently for each round, so that retrieval itself reflects the round's rhetorical purpose:
    - **Round 1:** the research question text is used directly.
    - **Round 2:** the research question is concatenated with the first 150 characters of the opponent's most recent response, biasing retrieval toward material relevant to the opponent's claims.
    - **Round 3:** the query is reformulated to emphasise evidence-bearing content, using the template *"specific example story evidence anecdote data about [topic]"*.

2. **Retrieve from the production database** with a metadata boost (described below), returning the top candidates by adjusted distance. When a research question filter is active, only chunks annotated for that research question are considered.
3. **Retrieve from the full database** without any boost, and add these results to the candidate pool.
4. **Remove duplicate chunks** across the two databases.
5. **Remove chunks already used in earlier rounds of the same debate**, so that each turn surfaces new material.
6. **Return the top five candidates** by adjusted distance.

Metadata boost. Candidates from the production database have their raw cosine distance reduced by up to 0.50 based on SCALE annotations: −0.35 if the chunk's primary category matches the round's assigned category, an additional −0.20 if the secondary category matches, and a further −0.15 if the chunk is rated quotable. The boost values were chosen so that a primary-category match outweighs surface semantic similarity, while quotability acts as a tiebreaker between rhetorically equivalent candidates.

A separate evidence retrieval method exists for cross-speaker evidence injection (described in Section 4.4.5). This method specifically targets Rhetorical Evidence chunks — those with a primary or secondary code of 3 — and applies a stricter semantic distance threshold of 0.75 for chunks from the full database. Evidence results are capped at four chunks per turn.

If the faithfulness filter (Section 4.4.4) returns a NEED_MORE_CHUNKS verdict, a re-retrieval method queries both databases using the filter's suggested search terms, prioritises the full database for depth, and applies SCALE boosts to any coded chunks found. The total number of chunks available to a single turn is capped at 20 to prevent prompt inflation.

4.4.3 Prompt Construction and Response Generation

Each bot's prompt is assembled from up to seven components, layered in a fixed order:

1. **Context block** — the concatenated text of all retrieved source chunks for the current turn.
2. **Evidence block** — any cross-speaker evidence chunks injected by the moderator from the opponent's corpus.

3. **Opponent's most recent response**, preceded by an instruction to respond to it directly.
4. **Moderator's guidance** for the current round, containing specific suggestions about claims to address and angles to explore.
5. **Filter feedback**, present only when a previous generation attempt failed quality evaluation — for example, "Your previous response contained claims not supported by your source material."
6. **Anti-repetition block** — the bot's own previous responses (truncated to 200 characters each) with an instruction to avoid repeating those points.
7. **Round-specific rhetorical instruction** corresponding to the current round's SCALE category.

The system prompt establishes the bot's persona and analytical framing. It identifies the bot as assisting with academic research on masculinity discourse for a bachelor's thesis at the University of Tartu, instructs it to simulate how the specified speaker would argue based only on the provided source material, and directs it to remain faithful to the speaker's documented views and rhetorical style. The prompt explicitly prohibits disclaimers, character breaks, and refusals, framing the task as scholarly analysis rather than role-play.

Response generation uses Claude Sonnet [27] with a maximum output length of 650 tokens and a temperature of 0.7. The target response length communicated in the prompt is 250 to 350 words. The prompt also prohibits section headers and structural markers, requiring the response to read as continuous argumentation rather than a formatted report. If the model refuses to generate a response, the system raises an error; truncated responses are accepted with a warning.

4.4.4 Faithfulness Filter

Every generated response passes through a four-criterion quality gate before publication. The filter uses a separate language model — Claude Haiku [28] at temperature 0.2 — to evaluate the response against the source chunks that were available to the bot during generation.

Each criterion is scored on a 1-to-5 scale. *Faithfulness* (threshold 4) evaluates whether every substantive claim in the response traces back to a source chunk, distinguishing genuine hallucinations — fabricated claims, false quotes, or character breaks contradicting the speaker's documented worldview — from acceptable framing such as transitional language or inferences

clearly implied by the chunks; as a hard override, any hallucinated claim caps faithfulness at 3, forcing a retry. *Engagement* (threshold 4) evaluates whether the response directly addresses the opponent's specific arguments, with a hard cap at 4 if the judge identifies four or more missed opponent points. *Specificity* (threshold 4) evaluates whether the response uses concrete details, examples, or phrasing drawn from the source chunks rather than generic text. *Novelty* (threshold 3, applied only in Rounds 2 and 3) evaluates whether the response introduces new arguments rather than recycling material from previous rounds; thematic continuity — returning to the speaker's core frameworks across rounds — is explicitly not penalised, with only the reuse of specific examples, stories, or phrases counted as recycling, and a hard cap at 3 if the judge identifies five or more recycled points.

The filter returns one of three verdicts. PASS indicates that the response meets all thresholds and is published. REGENERATE indicates that the response failed on faithfulness, specificity, or novelty; the same source chunks are retained, but the bot is re-prompted with specific feedback identifying what went wrong. NEED_MORE_CHUNKS indicates that the response failed on engagement, typically because the available chunks did not contain material relevant to the opponent's arguments; the retrieval system fetches additional chunks using search queries suggested by the filter before re-generating. A maximum of two retries is permitted. If all three attempts fail, the attempt with the highest combined score is published rather than discarding the turn entirely.

4.4.5 Moderation and Evidence Injection

The moderation layer consists of two components: a base moderator that guides the debate's progression, and an evidence-aware extension that facilitates cross-speaker evidence injection.

The base moderator performs three functions. After both bots have responded in a given round, the moderator analyses the round by identifying points of agreement, points of disagreement, the quality of arguments on both sides, and a suggested direction for the next round. This analysis uses Claude Haiku [28] at temperature 0.5. Before Rounds 2 and 3, the moderator generates concrete guidance for both bots using Claude Sonnet [27] at temperature 0.6. This guidance receives the full labelled transcript of the debate so far and must identify two to three specific claims from the previous round that the opposing debater should address, name sub-topics or angles not yet covered, call out explicit repetitions by naming the repeated point and the rounds in which it appeared, and avoid repeating its own previous guidance. After the final round, the moderator produces a synthesis of the entire debate, identifying the main themes

explored, areas of convergence, irreconcilable differences, the most compelling arguments from each side, and overall insights. This synthesis uses Claude Haiku at temperature 0.6 with a longer output allowance of 2,000 tokens.

The evidence-aware extension adds a cross-speaker evidence injection mechanism. After each round except the last, it identifies one to three claims per speaker that are central to the speaker's argument, specific enough that the opponent could find counter-evidence, and potentially challengeable with evidence or alternative framing. This identification step uses Claude Haiku at temperature 0.3 and produces a structured list of claim strings per speaker. Each speaker's identified claims are then used to retrieve Rhetorical Evidence chunks from the opposing speaker's corpus — Bot 1's claims trigger evidence retrieval from Bot 2's database, and vice versa. The retrieved evidence chunks are injected into the opponent's next-round prompt as supplementary context, enabling bots to counter specific claims with material from their own speaker's documented positions rather than generating unsupported rebuttals.

The moderation layer, together with the rest of the debate pipeline, is fully logged. For each debate the system preserves the full transcript and synthesis, every filter evaluation attempt with its scores and verdicts, and the per-round evidence identification and retrieval results.

5. Results and Analysis

This chapter reports the outcomes of the production pipeline described in Chapter 4 and analyses what those outcomes reveal about the masculinity discourse of the fourteen speakers in the corpus. The analysis is organised around three implicit research questions. First, can the SCALE multi-agent annotation framework reliably code masculinity discourse into the three rhetorical categories defined by the codebook? Second, does the retrieval-augmented debate system produce outputs that are faithful to each speaker's documented positions and that surface genuine ideological contrasts? Third, what does the resulting analysis reveal about the structure of contemporary masculinity discourse?

The first two questions concern the reliability of the analytical apparatus and must be answered before any substantive claims based on its outputs can be defended. The third question is the substantive contribution of the thesis and depends on the first two for its credibility. Section 5.1 reports the annotation reliability results and the speaker-level rhetorical patterns they reveal. Section 5.2 reports the debate system's quality-control performance across the production run, including the calibration choices that shaped that run. Section 5.3 examines the discursive contrasts and convergences that the system surfaces across the five research questions.

5.1 Reliability of Multi-Agent Annotation

The SCALE annotation pipeline produced consensus codes for the full production corpus of 7,000 chunks, distributed across 14 speakers and 5 research questions with 100 chunks per speaker–question combination. Each chunk received a primary code, a secondary code, and a quotability rating, derived through majority vote across three persona-conditioned LLM agents (Section 4.3.2) following up to two rounds of structured discussion (Section 4.3.3). This section reports four aspects of the annotation outcome: the inter-agent agreement that establishes the reliability of these codes, the convergence dynamics of the discussion phase, the typology of texts that resist consensus, and the distribution of rhetorical categories across speakers.

5.1.1 Inter-Agent Agreement

The agreement analysis was computed on the consensus-free initial coding pass: each agent's pre-discussion label, against every other agent's pre-discussion label, on the same 500 chunks per speaker (100 chunks × 5 research questions). Cohen's κ [32] — the standard pairwise agreement measure used in human inter-annotator reliability studies — was computed for each

pairwise combination of agents on each annotation dimension. Krippendorff's α [33] was computed across all three agents jointly. Table 2 reports the means and standard deviations across the 14 retained speakers.

Table 2. Inter-agent agreement on the initial (pre-discussion) SCALE coding pass, averaged across 14 speakers. Each cell reports mean ± SD; n = 500 paired codings per speaker.

Dimension	κ Sterling–Hayes	κ Sterling–Vance	κ Hayes–Vance	α (three-way)
Primary code	0.601 ± 0.069	0.595 ± 0.066	0.601 ± 0.056	0.596 ± 0.060
Secondary code	0.363 ± 0.051	0.364 ± 0.065	0.345 ± 0.054	0.355 ± 0.052
Quotability	0.643 ± 0.060	0.645 ± 0.061	0.621 ± 0.061	0.636 ± 0.052

Interpreted against the Landis and Koch [34] benchmarks for nominal agreement — fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) — primary-code agreement sits at the upper edge of the moderate band, quotability crosses into the substantial band, and the secondary code is fair. The Landis and Koch thresholds were developed for human raters applying mutually exclusive medical categories and their direct transfer to LLM coding of inherently overlapping rhetorical functions is not straightforward; the levels reported here are nonetheless comparable to those Gilardi, Alizadeh and Kubli [24] report for GPT-3.5 against trained human annotators on stance and topic-coding tasks ($\alpha \approx 0.50$–0.60).

The substantive pattern is the asymmetry across dimensions. Agreement on the primary code falls at the boundary between moderate and substantial, while the secondary code lags behind by roughly 0.24 α. This gap is partly intrinsic to the task: the secondary code asks coders to identify whether a faint subordinate rhetorical function is present alongside the dominant one — a judgment underdetermined on chunks of approximately 150 words and one that competent

human coders would be expected to find similarly difficult. The agents were also deliberately primed toward different categories (Section 4.3.2), which raises the baseline disagreement when the secondary code requires identifying a function the agent is not specialised in. Quotability, by contrast, asks a more constrained perceptual question — does the chunk contain a memorable self-contained formulation — and the agents converge on it more readily even though the agreement gain over chance is only modestly greater than the gain on the primary code.

One per-speaker pattern bears flagging because it returns in Section 5.1.4: Andrew Tate's corpus yields the highest agreement on both coding dimensions (primary-code $\alpha = 0.719$, secondary-code $\alpha = 0.504$), with quotability ($\alpha = 0.632$) also at the upper end though surpassed by ContraPoints ($\alpha = 0.755$), reflecting the stylistically distinctive character of his discourse, while the three lowest primary-code alphas all belong to Psychology-orientation speakers — Gottman ($\alpha = 0.489$), Perel ($\alpha = 0.495$), and Dr. K ($\alpha = 0.553$) — whose clinical-relational discourse blends prescription, anecdote, and reframing in single passages.

5.1.2 Test-Retest Reliability across Annotation Runs

A complementary question to within-run inter-agent agreement is whether SCALE produces consistent codes when run twice on the same text — its *test-retest reliability*. One speaker's video file was inadvertently annotated twice in the production run under different speaker identities (Section 4.3.6), with the same model, codebook, prompt templates, and chunking, but independent agent samples. The result is a paired dataset of 495 chunks of identical text — 99 per research question — annotated twice. The speaker in question is among those with mid-range within-run agreement, making the sample neither a best- nor worst-case comparison.

Table 3 reports Cohen's κ and percent agreement per research question and pooled.

Table 3. Test-retest reliability of the SCALE pipeline across two independent annotation runs on identical text (n = 99 paired chunks per RQ; n = 495 pooled).

Research question	n	Primary κ	Primary %	Secondary κ	Secondary %	Quotability κ	Quotability %

RQ1	99	0.498	76.8	0.237	44.4	0.626	86.9
RQ2	99	0.390	76.8	0.249	49.5	0.435	82.8
RQ3	99	0.612	80.8	0.389	55.6	0.558	89.9
RQ4	100	0.556	80.0	0.256	48.0	0.644	84.0
RQ5	98	0.497	73.5	0.339	52.0	0.529	81.6
Pooled	495	0.524	77.6	0.301	49.9	0.578	85.1

Two findings carry into the rest of the chapter. First, the within-run asymmetry between primary and secondary code reliability is reproduced across independent runs: the secondary code is consistently the least stable dimension. The structural ambiguity discussed in Section 5.1.1 is therefore a property of the task rather than an artifact of any single run. Second, primary-code agreement varies substantially across research questions — from $\kappa = 0.39$ on RQ2 to $\kappa = 0.61$ on RQ3 — indicating that what counts as the dominant rhetorical function depends partly on the analytical frame the question imposes on the text, and that the agents are sensitive to this dependency. Quotability, despite being binary, emerges as the most stable dimension across runs, suggesting that the judgment of whether a chunk contains a memorable self-contained formulation is more robust to LLM sampling variation than the judgment of which rhetorical function the chunk performs.

A pattern in the disagreement structure is worth flagging for Section 5.1.3: across all five research questions, the two runs agree closely on Provocation but trade Core Dogma and Rhetorical Evidence assignments back and forth. The Core Dogma ↔ Rhetorical Evidence

boundary is therefore the principal axis along which the framework is unstable, both across runs and — as Section 5.1.3 will show — across agents within a single run.

5.1.3 Boundary Cases and the Limits of Codebook Discrimination

The annotation pipeline reaches unanimous three-agent consensus across all three dimensions — primary code, secondary code, and quotability — on 4,045 of 7,000 chunks (57.8%); the remaining 2,955 chunks (42.2%) retain at least one dimension of disagreement after the two permitted discussion rounds. The SCALE judge evaluates these dimensions holistically, so a chunk is flagged as unresolved if any single dimension fails to converge. Decomposing the unresolved chunks by which dimension remains contested is therefore more diagnostic than the aggregate figure suggests. Table 4 reports the distribution.

Table 4. Decomposition of the 2,955 chunks that did not reach full three-dimensional unanimity after discussion. *Primary-unanimous* rows are chunks on which all three agents agreed on the dominant rhetorical function but disagreed on secondary code, quotability, or both. *Primary-code axis disputes* are 2-vs-1 splits confined to a single pair of primary codes.

Disagreement structure	n	% of unresolved	% of all chunks
Primary-unanimous: secondary code only	1,592	53.9	22.7
Primary-unanimous: quotability only	539	18.2	7.7
Primary-unanimous: secondary and quotability	123	4.2	1.8
Primary-code axis dispute: Core Dogma ↔ Rhetorical Evidence	439	14.9	6.3

Primary-code axis dispute: Core Dogma ↔ Provocation	188	6.4	2.7
Primary-code axis dispute: Provocation ↔ Rhetorical Evidence	74	2.5	1.1
Total	2,955	100.0	42.2

Two findings follow from the decomposition.

First, primary-code unanimity is substantially higher than the all-dimensions-unanimous figure implies. Across the full corpus, 6,299 of 7,000 chunks (90.0%) reach three-agent consensus on the primary rhetorical function after discussion, and only 701 chunks (10.0%) retain any primary-code disagreement. Of the 2,955 chunks classified as unresolved by the judge, 2,254 (76.3% of unresolved, 32.2% of all chunks) actually agree on the dominant rhetorical function and diverge only on whether a subordinate function is present or whether the chunk is quotable. Three-way primary-code splits — chunks for which all three agents assign different primary codes — are essentially absent, with only one chunk in the entire corpus exhibiting this pattern after discussion. The codebook discriminates well at the level it was primarily designed to address; instability is concentrated on the judgments that the analysis itself treats as more peripheral, consistent with the dimensional asymmetry reported in Section 5.1.1 and reproduced across independent runs in Section 5.1.2.

Second, among the 701 chunks with genuine primary-code disagreement, the Core Dogma ↔ Rhetorical Evidence boundary dominates, accounting for 62.6% of axis disputes. The same boundary emerged as the principal axis of test-retest disagreement in Section 5.1.2, suggesting that the difficulty is a property of the codebook–corpus combination rather than of any specific annotation attempt. A chunk in which a speaker articulates a fundamental belief by appealing to a personal anecdote occupies both categories simultaneously, and the disagreement persists through discussion because both readings are textually defensible. The dual-coding scheme described in Section 4.3.1 partially absorbs this ambiguity through the secondary code, but the assignment of primary versus secondary remains contested for these chunks.

The residual disagreement pattern after discussion is therefore interpretable rather than chaotic. The persona-aligned design described in Section 4.3.2 was intended to produce structured disagreement on chunks carrying faint signals of multiple rhetorical functions, and what survives discussion conforms to that expectation: minor differences over subordinate codes on chunks where the dominant function is clear, plus a smaller set of two-way primary disagreements on chunks where the categories themselves blur. SCALE's discussion mechanism removes the random three-way disagreements while preserving the substantive boundary disputes, exposing genuine codebook ambiguity rather than concealing it behind a single annotator's choice — one of the methodological advantages of multi-agent annotation over single-pass labelling [4].

5.1.4 Rhetorical Fingerprints across Speakers

Aggregating primary codes across all five research questions yields a per-speaker rhetorical distribution that functions as a discursive fingerprint. Table 5 reports the orientation-level means.

Table 5. Mean primary-code distribution per discursive orientation (mean ± SD across speakers in the orientation).

Orientation	n	Core Dogma	Provocation	Rhetorical Evidence
Radical	2	76.9 ± 1.3	10.1 ± 1.3	13.0 ± 0.0
Discipline	4	58.5 ± 9.3	7.4 ± 1.6	34.0 ± 10.7
Psychology	4	57.6 ± 8.4	14.2 ± 5.6	28.1 ± 11.4
Social Criticism	4	50.5 ± 10.9	18.0 ± 11.9	31.5 ± 5.6

The a priori four-orientation grouping is partially reflected in the empirical fingerprints. The Radical orientation shows the tightest within-group cohesion, with Tate and Fresh & Fit producing nearly identical profiles. The Discipline orientation is internally coherent on Provocation — all four speakers fall between 5% and 10% — but spreads widely on the Core Dogma / Rhetorical Evidence balance. The Psychology and Social Criticism orientations show the highest internal variance, particularly on Provocation: within Social Criticism, rates span a 30-point range from Armstrong (3.8%) to ContraPoints (34.6%), two speakers whose a priori grouping rested on a shared critical stance toward traditional masculinity rather than any shared rhetorical mode.

Cross-orientation affinities emerge that the a priori grouping does not predict. An evidence-heavy cluster, for example, links Huberman (Discipline), Gottman (Psychology), and Brown (Social Criticism), whose discourse is structured around concrete examples and case studies rather than declarative assertion — a methodological commitment that crosses ideological lines. Such cross-cutting patterns are visible throughout the fingerprint table and indicate that *how* speakers make claims operates partly independently of *what* they believe.

A counter-intuitive finding concerns the location of Provocation in the corpus. The Radical speakers — whose discourse is widely understood as provocative — do not lead on the Provocation code: Tate (11.4%) and Fresh & Fit (8.8%) record lower rates than Peterson, Dr. K, Destiny, and far lower than ContraPoints (34.6%). The Radical fingerprint is overwhelmingly Core Dogma. This is consistent with the codebook's instruction to classify by rhetorical function rather than tone (Section 4.3.1) and with the boundary analysis in Section 5.1.3: inflammatory delivery is most often coded as Core Dogma when it is constitutive of the belief being expressed. Provocation in this codebook indexes argumentative posture rather than ideological extremity, and the fingerprints make the distinction visible.

The fingerprint distribution has direct implications for the round–category mapping used by the debate system (Section 4.4.1). Round 1 (Core Dogma) is well-served across the corpus, and Round 3 (Rhetorical Evidence) is adequately supplied for most speakers. Round 2 (Provocation) is the most precarious: ten of the fourteen speakers have fewer than 12% of chunks coded as Provocation, and three — Armstrong (3.8%), Huberman (5.2%), Willink (6.8%) — fall below 7%, with Gottman a fourth case at exactly 7.0%. For these speakers, retrieval must rely on secondary-code matches and full-database fallback rather than primary-Provocation chunks. This empirically grounds the chunk-starvation phenomenon, motivates

the two-tier corpus architecture as a necessary design choice, and helps explain the corpus-level filter effects reported in Section 5.2.1.

5.2 Validity of the Debate System

The validity question for the debate system is whether its responses are faithful to each speaker's documented positions and whether contrasts between speakers survive into the generated text rather than being smoothed away by the underlying model. Two sources of evidence are reported here: the production behaviour of the four-criterion faithfulness filter described in Section 4.4.4, and a manual read-through of debate turns conducted by the author. A more rigorous human-validation study is left to future work; the implications are discussed in the Limitations chapter.

5.2.1 Filter Performance across the Production Run

Across the full production grid of 2,730 turns, the faithfulness filter passed 88.5 percent of responses on the first attempt. The remaining 11.5 percent entered the retry mechanism, which produced an acceptable response on 90.8 percent of retried turns; the residual 1.1 percent of all turns exhausted three attempts and were published under the best-of-three rule (Section 4.4.4). Failures were dominated by the faithfulness criterion: 88 percent of failed first attempts had a faithfulness violation flagged by the judge, consistent with the filter's design as primarily a grounding gate. Pass rates declined monotonically across rounds — 94 percent in Round 1, 89 percent in Round 2, 82 percent in Round 3 — reflecting the increasing demands the round-category mapping places on the speaker's annotated subset and the cumulative chunk-exclusion that constrains retrieval in later rounds.

A stricter filter configuration was tested by re-evaluating the production data against tighter thresholds; raising the novelty threshold by one point would have rejected roughly 70 percent of Round 2 and Round 3 first attempts. At that rate the retry budget is exhausted for a substantial fraction of debates, multiplying compute cost without improving the floor of the worst outputs, since those are also the responses for which retry cannot help. The production thresholds were therefore set at the level that detected grounding violations and severe recycling while keeping cost within the constraints of a bachelor's thesis.

5.2.2 Manual Read-through

A purposive sample of debate turns spanning the three rounds and the four discursive orientations was read manually by the author against the corresponding source chunks. Passed

responses remained recognisably in the speaker's voice, drew on transcript material rather than generic phrasing, and made claims defensible against the chunks the bot was given. The recycling the filter detects but tolerates was predominantly thematic continuity — speakers returning to their core frameworks across rounds with varied supporting material — rather than verbatim repetition, which the novelty rubric is explicitly designed to permit. Within the author's own knowledge of the speakers' published positions, the passed responses also captured ideological differences between speakers rather than collapsing toward a single model voice. Section 5.3 builds on this observation by analysing the full production grid quantitatively: if the system had smoothed speakers toward a single model voice, the position-similarity and convergence matrices reported there would show no structure beyond what random pair assignments would produce, and the substantive findings of that section would not be recoverable.

5.3 Findings on Masculinity Discourse

The four discursive orientations defined in Section 3.1 were constructed on the assumption that contemporary masculinity discourse is organised by ideological camp — Red Pill traditionalists against feminist critics, Stoic disciplinarians against relational therapists. The results below show that this assumption does not survive examination of how the speakers actually argue. What organises the corpus is not ideological position but rhetorical register and analytical level — how speakers make claims, rather than what they believe. Section 5.3.1 reports that the a priori four-orientation grouping is not recoverable from the data, using two complementary measures across the 91 speaker pairs: position similarity, computed from speakers' first-principles statements, and observed convergence, computed from the moderator-generated syntheses. Section 5.3.2 examines the relationship between these two measures, finds it weak (Pearson $r = 0.21$), and analyses the pairs where the two diverge most sharply — within-orientation rivals contesting the same conceptual ground, and cross-orientation speakers finding common ground when they share an analytical register.

5.3.1 The A Priori Orientations Do Not Emerge From the Data

The 14 speakers in this corpus do not group into the four ideological camps assigned to them. Two complementary measures were computed to test this. *Position similarity* measures how alike speakers' first-principles statements are: each speaker's most representative Round 1 response per research question was embedded with all-MiniLM-L6-v2, and the five resulting vectors were combined into a single per-speaker representation. *Observed convergence*

measures how often speakers found agreement in debate: for each of the 455 debates, the moderator-generated synthesis reports "Areas of Convergence" and "Irreconcilable Differences" as bulleted lists, and the ratio of agreement bullets to total bullets gives a per-debate convergence score, averaged across the five debates each pair participated in to produce a 14×14 matrix. The synthesis itself is generated by the debate pipeline and inherits the validity caveats discussed in Section 5.2; it is used here as a coarse-grained relational measure across the production grid rather than as a precise account of any individual debate's outcome.

Both matrices yield the same conclusion. Position similarity is narrowly distributed across all 91 speaker pairs (mean cosine similarity 0.66, range 0.54 to 0.73), and within-orientation mean similarity (0.66) is essentially identical to between-orientation mean similarity (0.66) — though the narrow range partly reflects the shared topic of masculinity, which guarantees baseline vector similarity regardless of stance. Agglomerative clustering applied to either matrix produced no groupings aligned with the a priori four-orientation labels, and the same null result held when the analysis was repeated separately on each research question: the most-similar speaker pair changed from question to question (Huberman × Destiny on RQ1, Fresh & Fit × Willink on RQ2, Tate × Destiny on RQ3, Tate × Peterson on RQ4, Gottman × Brown on RQ5), with no pair occupying the top position on more than one question and no stable orientation-aligned cluster recovered at any value of k.

The substantive finding is twofold. First, the popular framing of masculinity discourse as a contest between ideological camps is not reflected in what individual speakers actually say at the level of individual research questions: speakers who disagree on what masculinity should be may agree on what is wrong with men today, and vice versa. Second, the four-orientation grouping survives only as a coarse descriptive convention. The orientation labels remain useful for high-level characterisation of each group's claims (see Section 3.1), but they are an organising convention applied by the analyst, not a structural property of the data. Section 5.3.2 examines what *does* organise the discourse, when the apparent absence of ideological structure is replaced with a different distinction: between what speakers say and what they end up agreeing on in debate.

5.3.2 Position Similarity and Observed Convergence Are Different Phenomena

The two matrices introduced in Section 5.3.1 measure different things, and they are only weakly related. Across the 91 speaker pairs, position similarity correlates with observed

convergence at Pearson r = 0.21 (p = 0.05), explaining approximately four percent of the variance. The direction is the one a naive reader would expect — speakers with more similar stated positions converge slightly more often in debate — but the effect is too small to license predictions from ideological grouping to debate outcome. The substantive point holds: ideological similarity does not meaningfully predict productive dialogue, and ideological distance does not preclude it. The most informative cases are the pairs where these two measures diverge sharply.

Table 6. Pairs whose observed convergence diverges most sharply from their position similarity. Mismatch is computed as position similarity minus observed convergence. Positive values indicate pairs that should agree based on their stated positions but did not in debate; negative values indicate the opposite.

Pair	Position similarity	Observed convergence	Mismatch
Jocko Willink × Alison Armstrong	0.66	0.23	+0.43
Brené Brown × Alison Armstrong	0.65	0.23	+0.43
Jordan Peterson × Dr. K	0.69	0.29	+0.40
Brené Brown × Destiny	0.69	0.30	+0.39
Andrew Tate × Jordan Peterson	0.70	0.31	+0.39

Fresh & Fit × Andrew Huberman	0.68	1.00	−0.32
Andrew Tate × Fresh & Fit	0.68	1.00	−0.32
Esther Perel × Alison Armstrong	0.67	0.81	−0.14
ContraPoints × Destiny	0.69	0.77	−0.08
Fresh & Fit × Jordan Peterson	0.66	0.74	−0.08

The positive-mismatch pairs — speakers who hold similar positions but argue past each other — cluster around two recurring patterns, with one cross-cutting case.

The first pattern is *within-orientation rivalry*. Peterson × Dr. K is the clearest example: both Psychology speakers, both with high position similarity (0.69, among the top values in the corpus), but their debates produce the lowest within-Psychology convergence (0.29). The two occupy overlapping explanatory territory — both treat masculinity as a clinical-developmental object, both frame it in terms of competence and self-mastery — and they treat each other as competitors over the same conceptual ground rather than as allies. Brown × Destiny (similarity 0.69, convergence 0.30) shows the same pattern within Social Criticism. When speakers share both the topic and the analytical frame, the debate surfaces fine-grained disagreement over how to apply the shared frame, rather than the broad agreement one might expect.

The second pattern is *Armstrong as an outlier*. Alison Armstrong's positions embed at moderate-to-high similarity with both Brené Brown (0.65) and Jocko Willink (0.66), but her debates with each produce the corpus's two lowest convergence scores (both 0.23). The pattern is interpretable. Armstrong's framework treats traditional gendered complementarity as restorative — men are misrecognised "magnificent beings" whose qualities have been mistakenly relabelled as problems. Brown's framework treats traditional masculinity as a structure to be contested through vulnerability practice. Willink's treats masculinity as a

discipline of voluntary hardship. All three operate in a register that takes masculine struggle and masculine virtue as serious analytical objects, which is what produces the surface position similarity. But the prescriptions are incompatible: Armstrong reads the same evidence Brown and Willink read and arrives at the opposite conclusion. The Brown × Armstrong case is particularly instructive — both are nominally in the Social Criticism orientation on the basis that both contest the radical-masculinity project, but Brown contests it from a feminist-vulnerability framework and Armstrong contests it from an essentialist-complementarist framework. The orientation label predicts coalition; the actual debate produces the corpus's lowest convergence.

The remaining positive-mismatch pair, Tate × Peterson (similarity 0.70, convergence 0.31), fits neither pattern. Both speakers articulate hierarchical worldviews structured around competence, responsibility, and the legitimacy of male authority, which is what produces the high position similarity despite their cross-orientation status. Their low convergence reflects substantive disagreement over the *grounds* on which hierarchy is legitimated — Tate's framing rests on dominance and economic success, Peterson's on psychological development and Jungian archetypes — and this disagreement surfaces sharply in debate even though the structural commitment is shared. The pair functions as a third pattern in its own right: cross-orientation speakers whose underlying organisational principle aligns but whose justifications for that principle do not.

The negative-mismatch pairs — speakers who hold different positions but find agreement in debate — cluster around a different mechanism. Tate × Fresh & Fit (similarity 0.68, convergence 1.00) is the only intra-Radical pair in the corpus; the perfect convergence reflects that the two speakers reference each other directly and operate within the same content network, so it is best understood as an artefact of network adjacency rather than a finding about ideology. More analytically interesting are the cross-orientation cases. Fresh & Fit × Huberman (convergence 1.00 against moderate similarity 0.68) produces no recorded disagreements despite the speakers' nominally opposed orientations; reading the underlying transcripts, the convergence emerges through a shared biological-evolutionary register, in which both speakers treat male physiology as a legitimate object of intervention and the moderator's synthesis surfaces this shared framing as agreement even when specific prescriptions differ. Fresh & Fit × Peterson (similarity 0.66, convergence 0.74) shows the same mechanism on a different axis: both speakers organise their accounts of masculinity around hierarchy and competence, and the shared organisational register produces convergence in debate even though the substantive

positions — Red Pill dating market on one side, Jungian responsibility narrative on the other — would be expected to diverge. ContraPoints × Destiny (within Social Criticism, convergence 0.77 against similarity 0.69) shows the same pattern in a third register: both operate in a debate-rationalist mode that treats opponents' positions as objects of analysis rather than as moral claims, and the shared mode produces convergence across substantive disagreement. Perel × Armstrong (cross-orientation, convergence 0.81) is the most analytically interesting of these — two speakers from different a priori orientations who share a relational-developmental framework focused specifically on heterosexual partnership, finding common ground on partnership dynamics despite divergent positions on traditional gender roles.

The position-versus-convergence distinction is therefore a substantive feature of the corpus, not a methodological artifact. *What predicts productive dialogue is not ideological similarity but shared register, analytical level, and willingness to treat the opponent as an interlocutor.* This is consistent with the multi-agent debate literature reviewed in Section 2.2.3, particularly Smit et al. [22], who note that debate produces its strongest results when speakers can engage at compatible analytical levels rather than when they hold compatible positions. The system surfaces this distinction at scale.

6. Limitations

The findings of this thesis are qualified by limitations at two levels: the corpus and the debate system.

Corpus. The speaker sample was originally intended to include Christine Emba as a fifteenth speaker within the Social Criticism orientation, but her video corpus was excluded after transcript extraction failed to produce usable material. Three further videos were inadvertently scraped twice during corpus construction, affecting Fresh & Fit, Jordan Peterson, and Naval Ravikant; the duplicates represent approximately two percent of the 140 transcripts, and no single duplicate's chunks exceed roughly 0.7 percent of the affected speaker's total chunks, so the effect on retrieval is small relative to the category-aware distance boosts described in Section 4.4.2. Per-speaker corpus size varies almost threefold (Table 1), reflecting differences in speaking rate and interview length rather than thematic richness; for speakers with smaller corpora, retrieval overlap across the five research questions is correspondingly higher. Transcripts rely overwhelmingly on YouTube's automatic captioning system (129 of 140), with no manual correction of ASR errors. Downstream semantic-similarity retrieval is robust to surface-level transcription noise, but specific phrasings attributed to a speaker in the generated debate should not be treated as forensically exact quotations.

Debate system. No formal human-validation study of debate faithfulness was conducted. The four-criterion filter (Section 4.4.4) is itself an LLM-based judge, and the manual read-through reported in Section 5.2.2 was performed by the author alone on a purposive sample; stronger validity claims would require a stratified human annotation study and were left to future work. All three language models in the pipeline (Llama 3.3 70B for annotation, Claude Sonnet for generation, Claude Haiku for evaluation) are commercial-grade transformer LLMs sharing broad architectural and training-data characteristics, a setting known to produce central-tendency effects in multi-agent evaluation [23]. The observed-convergence scores reported in Section 5.3 should therefore be read as relational measures across the production grid rather than ground-truth indicators of how readily two speakers would actually agree in dialogue; a heterogeneous-model deployment, sketched in Chapter 7, is the natural next step.

7. Conclusion

This thesis set out to design, implement, and validate a computational pipeline for analysing the masculinity discourse of YouTube thought leaders, combining retrieval-augmented generation, multi-agent debate, and the SCALE collaborative annotation framework into a single integrated system. The validation question had two parts: whether the SCALE annotation produces consistent rhetorical codes across runs and agents (*reliability*), and whether the debate system generates responses faithful to each speaker's documented positions rather than to the underlying model's general assumptions about them (*validity*). Both had to be answered affirmatively before any substantive claim about the discourse itself could be defended.

The pipeline was built and run end-to-end over a corpus of 137 unique YouTube videos from 14 English-language speakers (approximately 3.02 million words and 23,776 retrieval-ready chunks). SCALE was adapted to classify transcript material into three rhetorical categories — Core Dogma, Provocation, and Rhetorical Evidence — using three persona-conditioned LLM agents and a structured discussion mechanism, and was integrated with the full corpus through a two-tier retrieval architecture that supports category-aware boosting without restricting retrieval to the annotated material alone. The debate system implements a fixed three-round structure mapped to the three rhetorical categories, with a four-criterion faithfulness filter and an evidence-aware moderator that injects cross-speaker counter-evidence between rounds. Annotation ran on the University of Tartu Rocket HPC cluster; debate generation and evaluation were distributed across three language models selected by cost–quality fit.

The reliability question was addressed across the production annotation grid of 7,000 chunks. Inter-agent agreement on the primary code reached the upper edge of the moderate band (Krippendorff's $\alpha = 0.596$), quotability the substantial band ($\alpha = 0.636$), and the secondary code the fair range ($\alpha = 0.355$). Test-retest reliability across two independent runs on identical text reproduced the same asymmetry, indicating that secondary-code instability is a property of the task rather than of any single run, and the residual primary-code disagreements concentrated on a single discriminative axis — Core Dogma versus Rhetorical Evidence — both within and across runs. These levels are comparable to those reported for LLM annotation on similarly subjective tasks, supporting the use of SCALE codes as input to downstream retrieval.

The validity question was addressed through the production behaviour of the faithfulness filter and a manual read-through. Across 2,730 generated turns the filter passed 88.5 percent of responses on first attempt and produced an acceptable response on 90.8 percent of retries, with

only 1.1 percent of turns exhausting the retry budget. Failures were dominated by faithfulness violations — the criterion the filter was primarily designed to enforce — and the manual sample confirmed that passed responses remained recognisably in each speaker's voice and drew on transcript material rather than collapsing toward a single model register.

With both validation questions addressed, the substantive analysis surfaced a finding that the a priori four-orientation grouping of speakers does not recover from the data, and that position similarity and observed convergence are only weakly correlated across the 91 speaker pairs. The cases in which the two diverge most sharply expose a consistent pattern: within-orientation rivals contest the same conceptual ground, while cross-orientation speakers find common ground when they share an analytical register. What predicts productive dialogue in this corpus is shared analytical level and rhetorical mode, not ideological proximity — a distinction that embedding-based or single-annotator methods are poorly placed to surface, and that the integrated pipeline makes visible.

The pipeline as built admits several natural extensions. The most consequential would be a heterogeneous-model deployment in which the two debate bots are served by language models with different training distributions, addressing the known ceiling effect that single-model architectures place on observable divergence. A human-validation study on a stratified sample of debate turns would strengthen the validity argument beyond what the filter and read-through can establish, and reactivating SCALE's human-intervention mechanism would allow expert adjudication of the Core Dogma versus Rhetorical Evidence boundary cases that the autonomous run cannot resolve. Within these constraints, the work demonstrates that retrieval-grounded, category-aware multi-agent debate provides a tractable computational route into discourse that is structured argumentatively rather than lexically — a class of analytical question to which embedding-based and single-annotator methods have so far had limited access.

References

[1] K. Lott, M. Murumaa-Mengel, and R. Marling, "Mainstreaming the manosphere: Discourses of contemporary masculinity among Estonian manfluencers," *Humanit. Soc. Sci. Commun.*, vol. 12, 2025.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[3] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 11733–11763.

[4] C. Zhao, Z. Tan, C.-W. Wong, X. Zhao, T. Chen, and H. Liu, "SCALE: Towards collaborative content analysis in social science with large language model agents and human intervention," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 8473–8503.

[5] R. W. Connell and J. W. Messerschmidt, "Hegemonic masculinity: Rethinking the concept," *Gend. Soc.*, vol. 19, no. 6, pp. 829–859, 2005.

[6] R. W. Connell, *Masculinities*. Cambridge: Polity Press, 1995.

[7] W. Khreich and J. Doughman, "Genderly: A data-centric gender bias detection system," *Complex Intell. Syst.*, vol. 11, p. 306, 2025.

[8] N. Bhatia and S. Bhatia, "Changes in gender stereotypes over time: A computational analysis," *Psychol. Women Q.*, vol. 45, no. 1, pp. 106–125, 2021.

[9] P. Kompatsiaris, "'Your Seed is Your Life Force': Masculinities of Spirituality and Entrepreneurialism in NoFap Lifestyle Gurus on YouTube," *J. Mens Stud.*, vol. 33, no. 3, pp. 527–548, 2025.

[10] E. G. Gram, B. Mintzes, T. Copp, R. Moynihan, A. Brown, P. Shih, and B. Nickel, "Selling masculinity – A qualitative analysis of gender representations in social media content about 'Low T'," *Soc. Sci. Med.*, vol. 393, art. no. 118903, 2025.

[11] S. Lucy, "Slippages in the application of hegemonic masculinity: A case study of incels," *Men and Masculinities*, vol. 27, no. 2, pp. 127–148, 2024.

[12] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?," *Comput. Linguist.*, vol. 50, no. 1, pp. 237–291, 2024.

[13] M. Shanahan, K. McDonell, and L. Reynolds, "Role play with large language models," *Nature*, vol. 623, no. 7987, pp. 493–498, 2023.

[14] T. Hu and N. Collier, "Quantifying the persona effect in LLM simulations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 10289–10307.

[15] M. Cheng, T. Piccardi, and D. Yang, "CoMPosT: Characterizing and evaluating caricature in LLM simulations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 10853–10875.

[16] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[17] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3784–3803.

[18] Q. Huang, S. Fu, X. Liu, W. Wang, T. Ko, Y. Zhang, and L. Tang, "Learning retrieval augmentation for personalized dialogue generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2523–2540.

[19] "Chroma: the open-source AI application database," GitHub. Accessed: May 10, 2026. [Online]. Available: https://github.com/chroma-core/chroma

[20] R. Qu, R. Tu, and F. S. Bao, "Is semantic chunking worth the computational cost?," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 2155–2177.

[21] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, "Encouraging divergent thinking in large language models through multi-agent debate," in

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17889–17904.

[22] A. P. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius, "Should we be going MAD? A look at multi-agent debate strategies for LLMs," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 45883–45905.

[23] H. Wu, Z. Li, and L. Li, "Can LLM agents really debate? A controlled study of multi-agent debate in logical reasoning," arXiv:2511.07784, 2025. [Online]. Available: https://arxiv.org/abs/2511.07784

[24] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Natl. Acad. Sci.*, vol. 120, no. 30, p. e2305016120, 2023.

[25] J. Depoix, *youtube-transcript-api*. GitHub, 2014. Accessed: Mar. 27, 2026. [Online]. Available: https://github.com/jdepoix/youtube-transcript-api

[26] A. Dubey, A. Jauhri, A. Pandey, *et al.*, "The Llama 3 herd of models," arXiv:2407.21783, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[27] Anthropic, "Claude Sonnet 4." Accessed: May 11, 2026. [Online]. Available: https://www.anthropic.com/claude/sonnet

[28] Anthropic, "Claude Haiku 4.5." Accessed: May 11, 2026. [Online]. Available: https://www.anthropic.com/claude/haiku

[29] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[30] University of Tartu HPC Centre, "Rocket cluster." Accessed: May 14, 2026. [Online]. Available: https://hpc.ut.ee/

[31] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023, pp. 611–626.

[32] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[33] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Sage, 2018.

[34] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

Appendices

Appendix A. Source Videos

The corpus comprises 137 unique YouTube videos across fourteen speakers. The list below is organised by the four discursive orientations defined in Section 3.1. URLs were valid at the time of access (March 2026).

A.1 Radical masculinity and “Red Pill” ideology

Andrew Tate

1. https://www.youtube.com/watch?v=lom1D8raoA4
2. https://www.youtube.com/watch?v=gnZmRWTSyLs
3. https://www.youtube.com/watch?v=2ELmm02-jCQ
4. https://www.youtube.com/watch?v=iv-C4CVGk28
5. https://www.youtube.com/watch?v=KMy8nwEWnvg
6. https://www.youtube.com/watch?v=HZ6FhDfa9UY
7. https://www.youtube.com/watch?v=1nvcZaD8vQk
8. https://www.youtube.com/watch?v=_-xkTfdKfnw
9. https://www.youtube.com/watch?v=88ibAvTtWLc
10. https://www.youtube.com/watch?v=m1S9oOgnGp0

Fresh & Fit

1. https://www.youtube.com/watch?v=HOycSS3sfQk
2. https://www.youtube.com/watch?v=aUQDrZy30G0
3. https://www.youtube.com/watch?v=Let0TXmvpdk
4. https://www.youtube.com/watch?v=gxuS6A0igCk
5. https://www.youtube.com/watch?v=yV0jdJCWqig
6. https://www.youtube.com/watch?v=OjfOkb203Mo

7. https://www.youtube.com/watch?v=kK3xuYuduiE
8. https://www.youtube.com/watch?v=Fvf-MMBlLjY
9. https://www.youtube.com/watch?v=fCcLJcyFbPg

A.2 Discipline, stoicism, and self-optimisation

Jocko Willink

1. https://www.youtube.com/watch?v=3eZJ_y68wsg
2. https://www.youtube.com/watch?v=Cu0jLtOlHeY
3. https://www.youtube.com/watch?v=4W64WGFy-Js
4. https://www.youtube.com/watch?v=NnKcquMobHQ
5. https://www.youtube.com/watch?v=YeJ4b6eTmq8
6. https://www.youtube.com/watch?v=HA4Bkybx1ps
7. https://www.youtube.com/watch?v=bL5RzI5LyVc
8. https://www.youtube.com/watch?v=__RAXBLt1iM
9. https://www.youtube.com/watch?v=DgNzZsGZ96Q
10. https://www.youtube.com/watch?v=nFYvmTWHhnc

Andrew Huberman

1. https://www.youtube.com/watch?v=jSqCL7Npln0
2. https://www.youtube.com/watch?v=spq8UKib3Zw
3. https://www.youtube.com/watch?v=gxHktBHVY5I
4. https://www.youtube.com/watch?v=SwQhKFMxmDY
5. https://www.youtube.com/watch?v=BoutTY8XHSc
6. https://www.youtube.com/watch?v=VSG9hY_t-rs
7. https://www.youtube.com/watch?v=gmtRU-IqPZk
8. https://www.youtube.com/watch?v=z-mJEZbHFLs

9. https://www.youtube.com/watch?v=31DMZLK_PPs
10. https://www.youtube.com/watch?v=gLJowTOkZVo

Ryan Holiday

1. https://www.youtube.com/watch?v=PafvhTSC4yE
2. https://www.youtube.com/watch?v=VKJNwcLKsl8
3. https://www.youtube.com/watch?v=vGXvQqurY-Y
4. https://www.youtube.com/watch?v=stnG5IGZc4k
5. https://www.youtube.com/watch?v=_PMesNbFuFQ
6. https://www.youtube.com/watch?v=CWvOKHNeLMI
7. https://www.youtube.com/watch?v=VjOCMnJVJa0
8. https://www.youtube.com/watch?v=IAMdQLyoK70
9. https://www.youtube.com/watch?v=bv4DlMnlQn4
10. https://www.youtube.com/watch?v=HvEg37B4DU4

Naval Ravikant

1. https://www.youtube.com/watch?v=KyfUysrNaco
2. https://www.youtube.com/watch?v=3qHkcs3kG44
3. https://www.youtube.com/watch?v=mGY2To_HW98
4. https://www.youtube.com/watch?v=FfWbcrObpUY
5. https://www.youtube.com/watch?v=Z1cz3MyjlBo
6. https://www.youtube.com/watch?v=0HiUWufjzuM
7. https://www.youtube.com/watch?v=10Wkb7hD0cU
8. https://www.youtube.com/watch?v=P-FnidFTBhQ
9. https://www.youtube.com/watch?v=wQGOYnWHnto

A.3 Psychology, relationships, and mental health

Jordan Peterson

1. https://www.youtube.com/watch?v=aMcjxSThD54
2. https://www.youtube.com/watch?v=yZYQpge1W5s
3. https://www.youtube.com/watch?v=Ng7EjFEMSp8
4. https://www.youtube.com/watch?v=BTv5feWd9dk
5. https://www.youtube.com/watch?v=7QRQjrsFnR4
6. https://www.youtube.com/watch?v=1ruwVc0t2c0
7. https://www.youtube.com/watch?v=QBEZhjnZTks
8. https://www.youtube.com/watch?v=6T7pUEZfgdI

Esther Perel

1. https://www.youtube.com/watch?v=uLB4qGYMO6s
2. https://www.youtube.com/watch?v=zoF_eUfzjwY
3. https://www.youtube.com/watch?v=ss1zseYy7J8
4. https://www.youtube.com/watch?v=r0dVjeBNANA
5. https://www.youtube.com/watch?v=ajneRM-ET1Q
6. https://www.youtube.com/watch?v=nTWXfo7narw
7. https://www.youtube.com/watch?v=wXoi2mcS79c
8. https://www.youtube.com/watch?v=S57343tPKHg
9. https://www.youtube.com/watch?v=XS38vqKDiEo
10. https://www.youtube.com/watch?v=r8tqC63ZLDw
11. https://www.youtube.com/watch?v=QCaFWrT0j-g

John Gottman

1. https://www.youtube.com/watch?v=VJd3fM4kgHs
2. https://www.youtube.com/watch?v=YbaCdTwHBLs

3. https://www.youtube.com/watch?v=LwSQ6r54xSw
4. https://www.youtube.com/watch?v=bkDQnslKzrE
5. https://www.youtube.com/watch?v=8YpnDao8TQs
6. https://www.youtube.com/watch?v=mS3bfCt0K88
7. https://www.youtube.com/watch?v=1g1X0heybWA
8. https://www.youtube.com/watch?v=JKj3lReSzJk
9. https://www.youtube.com/watch?v=Bdazx9VgP44
10. https://www.youtube.com/watch?v=ZQ_QKmuIjQA

Dr. K (HealthyGamerGG)

1. https://www.youtube.com/watch?v=B_5N_aDu3u0
2. https://www.youtube.com/watch?v=P1ALkQMfkjc
3. https://www.youtube.com/watch?v=4ZFo207xago
4. https://www.youtube.com/watch?v=PMmObJYglbo
5. https://www.youtube.com/watch?v=dm2qDrb3UVo
6. https://www.youtube.com/watch?v=dMFhqbfnIfQ
7. https://www.youtube.com/watch?v=u-XDOnSSUzI
8. https://www.youtube.com/watch?v=5-dc3tP1Mr4
9. https://www.youtube.com/watch?v=bbvz4j-oPrM
10. https://www.youtube.com/watch?v=CxAAiGz_TlY

A.4 Social criticism and alternative perspectives

Brené Brown

1. https://www.youtube.com/watch?v=jn86FtQMWHQ
2. https://www.youtube.com/watch?v=TbsRU-crgsc
3. https://www.youtube.com/watch?v=R1Zp9b6iV14

4. https://www.youtube.com/watch?v=vCpYHYmHHvY
5. https://www.youtube.com/watch?v=SM1ckkGwqZI
6. https://www.youtube.com/watch?v=NgmJinwZDgw
7. https://www.youtube.com/watch?v=Wh5SUF0gPWQ
8. https://www.youtube.com/watch?v=jroF3PH-PTs
9. https://www.youtube.com/watch?v=Z0Atlk2Qr8A
10. https://www.youtube.com/watch?v=jddNUhUL-uI

ContraPoints (Natalie Wynn)

1. https://www.youtube.com/watch?v=cKrxP44Gp_0
2. https://www.youtube.com/watch?v=V445G-ftJqY
3. https://www.youtube.com/watch?v=f7yco_wZvPk
4. https://www.youtube.com/watch?v=xedSSsQeGL0
5. https://www.youtube.com/watch?v=5AjeEoNQ5tw
6. https://www.youtube.com/watch?v=K5gI2RicywA
7. https://www.youtube.com/watch?v=teGyINp-qRs
8. https://www.youtube.com/watch?v=S1xxcKCGljY
9. https://www.youtube.com/watch?v=1pTPuoGjQsI
10. https://www.youtube.com/watch?v=4LqZdkkBDas

Destiny (Steven Bonnell)

1. https://www.youtube.com/watch?v=xqnPhSgCh28
2. https://www.youtube.com/watch?v=u4lYO98fYwo
3. https://www.youtube.com/watch?v=Eck9JwSeTI0
4. https://www.youtube.com/watch?v=tb4zpUAgyqw
5. https://www.youtube.com/watch?v=9-x9dGUc92Q

6. https://www.youtube.com/watch?v=jOqNAf1aHoY
7. https://www.youtube.com/watch?v=2IYuqfrMkMo
8. https://www.youtube.com/watch?v=Qw9Dcw-HtVo
9. https://www.youtube.com/watch?v=mxTxJHbRg0U
10. https://www.youtube.com/watch?v=KUHcveLBNI8

Alison Armstrong

1. https://www.youtube.com/watch?v=4hibdmZkOIE
2. https://www.youtube.com/watch?v=5j5tJNmse2I
3. https://www.youtube.com/watch?v=ZLUfYd7GikM
4. https://www.youtube.com/watch?v=UoPRWA9De5E
5. https://www.youtube.com/watch?v=hsZJYRhVmZw
6. https://www.youtube.com/watch?v=sjZ-F_pC548
7. https://www.youtube.com/watch?v=EzDy3xlZIo4
8. https://www.youtube.com/watch?v=rMDomNCVAjI
9. https://www.youtube.com/watch?v=Cy5G3lK9HJk
10. https://www.youtube.com/watch?v=RfyP_oPGmMI

Appendix B. SCALE Configuration

This appendix documents the configuration of the SCALE annotation pipeline as used in the production run described in Section 4.3.6. It contains the codebook (Section B.1), the three agent persona prompts (Section B.2), the prompt templates that drive the coding, discussion, judging, and codebook-update steps (Section B.3). All strings are reproduced verbatim from the version used to annotate the production corpus.

B.1 Codebook

The codebook below is the version used for the final production run. Codebook evolution was disabled (see Section 4.3.6), so this text was loaded once at the start of the run and was not modified across batches.

CODEBOOK: Masculinity Discourse Analysis

For each transcript chunk, provide THREE annotations:

A) PRIMARY RHETORICAL FUNCTION [Choose one: 1, 2, or 3]

B) SECONDARY RHETORICAL FUNCTION [Choose one: 1, 2, 3, or 0 if none]

C) QUOTABILITY [Choose: Q or NQ]

RHETORICAL FUNCTION CATEGORIES:

1. Core Dogma (Belief/Worldview): The text primarily asserts
a fundamental rule, philosophical truth, or definition about
masculinity, society, or human nature. This includes:

- Declarations of how men 'should' behave or what constitutes a 'real man'.

- Statements establishing hierarchy, stoicism, or duty.

- Examples: 'A man's value is derived from his ability to protect
and provide.', 'Meaning is found through taking on responsibility.'

2. Provocation (Attack/Polemic): The text primarily challenges
opposing viewpoints or reframes assumptions to destabilize
them. This includes:

- Insults directed at 'weakness', modern culture, or ideological opponents.

- Deliberately controversial statements designed to trigger outrage
or assert dominance.

- Examples: 'The people pushing for vulnerability just want to
emasculate you.', 'If you are a victim, you are useless.'

3. Rhetorical Evidence (Anecdote/Metaphor): The text primarily

uses personal stories, clinical examples, biological claims,

or metaphors to justify a previously established belief. This includes:

- Evolutionary psychology claims or references to animal behavior.

- Clinical anecdotes or personal success stories.

- Examples: 'Look at lobsters; their nervous systems are wired for

hierarchy.', 'In my practice, I saw hundreds of young men who were

completely lost.'

SECONDARY FUNCTION:

Assign a secondary code (1, 2, or 3) if the chunk clearly

serves a second rhetorical function beyond its primary one.

Assign 0 if the chunk serves only one function.

QUOTABILITY:

Q = Contains a memorable, self-contained phrase effective

as a direct quote in a debate context.

NQ = No standout quotable phrase.

Note: These chunks come from multiple speakers with different

rhetorical styles. The same category may manifest differently

across speakers. Code based on rhetorical function, not tone.

B.2 Agent Persona Prompts

Each of the three agents introduced in Section 4.3.2 was initialised with a fixed persona system prompt. The full text of each prompt is reproduced below.

Sterling — Political Sociologist.

You are Dr. Sterling, a 55-year-old Political Sociologist analyzing

modern ideological movements. Your expertise lies in identifying core

dogmas, power dynamics, and structural worldviews. You approach texts objectively, looking for fundamental belief statements regarding hierarchy, society, and human nature. You do not judge the morality of the text; you strictly identify its underlying ideological framework.

Hayes — Gender Studies and Clinical Psychology.

You are Dr. Hayes, a 42-year-old researcher in Gender Studies and Clinical Psychology. You analyze how gender roles, emotional norms, and societal expectations are constructed. You look for statements that define what it means to be a 'real man,' attitudes toward emotional vulnerability, and relational dynamics. You maintain a strictly academic and analytical tone.

Vance — Rhetoric and Persuasion.

You are Dr. Vance, a 38-year-old expert in Rhetoric and Persuasion. You analyze texts to identify debate tactics, framing, and provocative statements. You look for inflammatory language, polemics, metaphors, and how the speaker attacks opposing viewpoints or asserts dominance over the audience. You focus strictly on *how* the argument is constructed and weaponized.

B.3 Prompt Templates

The pipeline uses six prompt templates. The **coding**, **discussion**, and **judge** prompts drive every annotation. The **update** and **mediator** prompts implement codebook evolution, which was disabled for the production run (Section 4.3.6) but is retained in the configuration to document the framework capability. The **collaborative** and **directive** prompts implement the human-intervention mechanism described in Section 2.3, which was not invoked because the pipeline ran in fully autonomous mode (Section 4.3.3).

Coding prompt. Issued to each agent during the independent coding phase.

You are now coding text entries independently following the instructions:

1. Process each TEXT using the guidelines in the CODEBOOK. The CODEBOOK requires THREE annotations per text: (A) PRIMARY rhetorical function [1, 2, or 3], (B) SECONDARY rhetorical function [1, 2, 3, or 0 if none], and (C) QUOTABILITY [Q or NQ].

2. Base decisions solely on the CODEBOOK and PERSONA.

3. Act as a social scientist, providing a well-reasoned explanation for each of the three annotations. In your explanation, quote the exact sentence from the text that justifies your coding.

4. At the end of your response, state your final answers in this exact format, each on its own line:

PRIMARY: <1|2|3>

SECONDARY: <1|2|3|0>

QUOTABILITY: <Q|NQ>

Discussion prompt. Issued to each agent during the consensus-seeking discussion phase when the judge has detected disagreement.

For some TEXTs, other social scientists have provided different coding results and reasons. You are now conducting a discussion. Below are the responses from other social scientists. Use these responses carefully as additional guidance. You may accept or reject their opinions when updating your three annotations (PRIMARY, SECONDARY, QUOTABILITY). Keep your response under 200 words. State only your updated position and the single strongest reason for any change. At the end of your response,

state your final answers in this exact format, each on its own line:

PRIMARY: <1|2|3>

SECONDARY: <1|2|3|0>

QUOTABILITY: <Q|NQ>

Judge prompt. Issued to a separate judge agent that decides whether the three annotators have converged.

You are a judge. Several social scientists are working on a content

analysis task. Compare their responses and determine if they agree.

During CODING or DISCUSSION: each scientist provides annotations in

the format 'PRIMARY: X', 'SECONDARY: Y', 'QUOTABILITY: Z'. Respond

'same' ONLY if all three annotations match exactly across all

scientists. During CODEBOOK REVIEW: each scientist states whether

they accept or reject a proposed codebook. Respond 'same' if all

scientists take the same position (all accept OR all reject). Respond

'different' if any scientist's position differs from the others. Only

include the single word 'same' or 'different' in your response.

Update prompt (not invoked in the production run). Issued to each agent after a batch to propose codebook revisions.

Based on the coding and discussion results, you are now updating the

CODEBOOK. You may revise the CODEBOOK or keep it unchanged. Do not

change the CODEBOOK if it adequately fits the current examples. If you

make changes, output the UPDATED CODEBOOK; otherwise, output the

ORIGINAL CODEBOOK.

Criteria for a good CODEBOOK:

1. The CODEBOOK should cover all cases and patterns in the examples.

2. Each rule should be unique, with minimal overlap.

Guidelines for changes:

1. You may add examples or clarifications to the categories.

2. Do not change the existing fundamental categories.

Mediator prompt (not invoked in the production run). Issued to a mediator agent that synthesises multiple proposed codebook revisions into a single updated codebook.

You are a mediator. The social scientists have proposed updates to the

CODEBOOK. Synthesize their proposals into a single unified CODEBOOK.

You MUST preserve the fundamental structure: THREE annotations per text

(PRIMARY [1,2,3], SECONDARY [1,2,3,0], QUOTABILITY [Q,NQ]) using the

three rhetorical function categories (1=Core Dogma, 2=Provocation,

3=Rhetorical Evidence). You may ONLY add clarifying examples or refine

descriptions within existing categories. Do NOT add, remove, rename, or

replace any categories. Do NOT change the SECONDARY codes from 1,2,3,0

to anything else. Output only the content of the CODEBOOK.

Collaborative-mode intervention prompt (not invoked in the production run). Used when a human expert provides advisory input that the agent may accept or reject.

Another social scientist has provided advice on your response. Consider

this advice carefully. You may accept or reject it when updating your

answer.

Directive-mode intervention prompt (not invoked in the production run). Used when a human expert issues binding instructions that the agent must follow.

A human social scientist expert has issued instructions regarding your

response. You MUST follow these instructions when updating your answer.

Licence

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Kim Lilii Tamm**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose **Modelling Masculinity Ideals Using Machine Learning: Opinion Leader Discourse Analysis Through Debate**, mille juhendaja on **Krista Liin**, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;

3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;

4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kim Lilii Tamm

14.05.2026