

The General Social Survey



2024 GSS
(Cross-section
Study)

LLM Data File
Documentation



2024 GSS LLM Data File

Introduction

With the third release of MR #150, NORC is making available a supplementary data file as a CSV file that includes the human and LLM coding output of open-ended responses from HHTYPE and TOPPROB from the 2024 GSS follow-on questionnaire. As part of an experiment in the GSS Next follow-on study, two open-ended questions were asked of respondents covering two domains: the most important issues facing America today (TOPPROB) and respondents' household composition (HHTYPE). These questions differ in complexity, with one opinion-based and nuanced, and the other more factual and difficult to capture with standard question batteries. Respondents were asked to classify their open-ended responses into predefined GSS response categories, enabling direct comparison between the LLM coding versus respondent-provided classification. We test three LLMs across two prompting conditions to evaluate the output of LLM coding with traditional human coding. We include the coding output from this testing in a datafile for users interested in this research. We strongly encourage users to familiarize themselves with MR 150 before working with this data.

Table 1 details the variables available in this data file. It includes GSS respondent YEAR, ID, NEXTSTATUS HHTYPE1_NEXT, TOPPROB1_NEXT, TOPPROB1, MARITAL, CHILDS, AGE, SEX, BORN, and FAMGEN_NEXT; all of which are found in the GSS 2024 Individual Year Datafile and the GSS 1972-2024 Cumulative data file. The results of LLM coding research are additionally in the datafile, which include coding results across HHTYPE and TOPPROB open-ended data: human coding, and coding results from three LLMs (GPT-5-Mini, GPT-5, and Claude-Sonnet-4.5), and their LLM-produced confidence scores (denoted with suffix "_conf"), across two prompting conditions (denoted with "P1" and "P2"). Last, users will see some GSS reserve codes such as ".i: Inapplicable", ".d: Don't know", ".n: No answer", or ".s: Skipped on web". Table 1 reports the order of the data file, showing HHTYPE coding output followed by TOPPROB coding output. Please see Table 2 for the HHTYPE coding categories and Table 3 for the TOPPROB coding categories. Additionally, please see the Codebook in this package for more information on the variables, including variable labels and reserve codes. The data file does not include the verbatim responses to the open-ended question.

The data that are made available in this datafile are associated with the 2024 General Social Survey Next follow-on. They do not receive the same level of cleaning and harmonization as the contents of the GSS 1972-2024 Cumulative Datafile or other data releases (e.g., 2024 GSS Release 3). As such, users should expect potential inconsistencies and improbable values due to the experimental nature of this research. These can include values with unexpected decimals and values that are outside of the value range for that code frame. The GSS is releasing these data to support methodological research. Should users have specific questions on these data, they are encouraged to contact the GSS (gss@norc.org).

Table 1: Variable Availability

Variable	Response Options	Previous Variable	Next Variable
YEAR	2024	-	ID
ID	1-3309	YEAR	NEXTSTATUS
NEXTSTATUS	1	ID	HHTYPE1_NEXT
HHTYPE1_NEXT	1-11	NEXTSTATUS	HHTYPE_Expert
HHTYPE_Expert	1-11	HHTYPE1_NEXT	HHTYPE_P1_GPT5mini
HHTYPE_P1_GPT5mini	1-11	HHTYPE_Expert	HHTYPE_P1_GPT5mini_conf
HHTYPE_P1_GPT5mini_conf	0-1	HHTYPE_P1_GPT5mini	HHTYPE_P1_GPT5
HHTYPE_P1_GPT5	1-11	HHTYPE_P1_GPT5mini_conf	HHTYPE_P1_GPT5_conf
HHTYPE_P1_GPT5_conf	0-1	HHTYPE_P1_GPT5	HHTYPE_P1_Sonnet
HHTYPE_P1_Sonnet	1-11	HHTYPE_P1_GPT5_conf	HHTYPE_P1_Sonnet_conf
HHTYPE_P1_Sonnet_conf	0-1	HHTYPE_P1_Sonnet	FAMGEN_NEXT
FAMGEN_NEXT	1. 1 gen. 2. 2 gen., children 3. 2 gen., parents 4. 2 gen., grandchildren 5. 3+ gen.	HHTYPE_P1_Sonnet_conf	MARITAL
MARITAL	1. Married 2. Widowed 3. Divorced 4. Separated 5. Never married	FAMGEN_NEXT	CHILDS
CHILDS	0-8	MARITAL	AGE
AGE	18-89	CHILDS	SEX
SEX	1. Male	AGE	BORN

Variable	Response Options	Previous Variable	Next Variable
	2. Female		
BORN	1. Yes 2. No	SEX	HHTYPE_P2_GPT5mini
HHTYPE_P2_GPT5mini	1-11	BORN	HHTYPE_P2_GPT5mini_conf
HHTYPE_P2_GPT5mini_conf	0-1	HHTYPE_P2_GPT5mini	HHTYPE_P2_GPT5
HHTYPE_P2_GPT5	1-11	HHTYPE_P2_GPT5mini_conf	HHTYPE_P2_GPT5_conf
HHTYPE_P2_GPT5_conf	0-1	HHTYPE_P2_GPT5	HHTYPE_P2_Sonnet
HHTYPE_P2_Sonnet	1-11	HHTYPE_P2_GPT5_conf	HHTYPE_P2_Sonnet_conf
HHTYPE_P2_Sonnet_conf	0-1	HHTYPE_P2_Sonnet	TOPPROB1_NEXT
TOPPROB1_NEXT	1-9	HHTYPE_P2_Sonnet_conf	TOPPROB_Expert
TOPPROB_Expert	1-9	TOPPROB1_NEXT	TOPPROB_P1_GPT5mini
TOPPROB_P1_GPT5mini	1-9	TOPPROB_Expert	TOPPROB_P1_GPT5mini_conf
TOPPROB_P1_GPT5mini_conf	0-1	TOPPROB_P1_GPT5mini	TOPPROB_P1_GPT5
TOPPROB_P1_GPT5	1-9	TOPPROB_P1_GPT5mini_conf	TOPPROB_P1_GPT5_conf
TOPPROB_P1_GPT5_conf	0-1	TOPPROB_P1_GPT5	TOPPROB_P1_Sonnet
TOPPROB_P1_Sonnet	1-9	TOPPROB_P1_GPT5_conf	TOPPROB_P1_Sonnet_conf
TOPPROB_P1_Sonnet_conf	0-1	TOPPROB_P1_Sonnet	TOPPROB_P2_GPT5mini
TOPPROB_P2_GPT5mini	1-9	TOPPROB_P1_Sonnet_conf	TOPPROB_P2_GPT5mini_conf
TOPPROB_P2_GPT5mini_conf	0-1	TOPPROB_P2_GPT5mini	TOPPROB_P2_GPT5
TOPPROB_P2_GPT5	1-9	TOPPROB_P2_GPT5mini_conf	TOPPROB_P2_GPT5_conf
TOPPROB_P2_GPT5_conf	0-1	TOPPROB_P2_GPT5	TOPPROB_P2_Sonnet
TOPPROB_P2_Sonnet	1-9	TOPPROB_P2_GPT5_conf	TOPPROB_P2_Sonnet_conf
TOPPROB_P2_Sonnet_conf	0-1	TOPPROB_P2_Sonnet	-

Table 2: HHTYPE Coding Categories

Value	Category
1	Married couple, no children
2	Single parent
3	Other family, no children
4	Single adult
5	Cohabiting couple, no children
6	Non-family, no children
7	Married couple, children
8	Cohabiting couple, children
9	Other family relationship, children
10	Non-family relationship, children
11	None of these

Table 3: TOPPROB Coding Categories

Value	Category
1	Health care
2	Education
3	Crime
4	The environment
5	Immigration
6	The economy
7	Terrorism
8	Poverty
9	None of these