Precision Shapes Personality: The Hidden Cost of Quantization in Sub-Billion-LLMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Psychometric studies of language models are increasingly important given their growing use as human assistants and in therapeutic settings. Such applications are often deployed on edge devices with sub-billion parameter large language models (LLMs) operating under strict memory and latency constraints, where post-training quantization (PTQ) is standard. Yet little is known about whether numeric precision alters measured personality traits. In the current work, using a psychometric benchmark TRAIT, we evaluate five sub-1B LLMs across different precision settings. We find that 4-bit Normal Float (nf4) produces the largest shifts, int8 smaller ones, and 16-bit formats remain closest to native. Shifts concentrate in Extraversion, Conscientiousness, and Narcissism, while Openness and Machiavellianism are more stable. These results identify precision as a consequential, controllable variable that should be disclosed and audited when personality matters in deployment.

1 Introduction

Sub-Billion Parameter Large Language Models (LLMs) are increasingly deployed on phones, wearables, and other constrained devices, where post-training quantization (PTQ) reduces memory/latency and enables real-time inference [1–3]. If precision alone alters measured traits, aggressive 4-bit settings could be risky in practice. Even when prompts and decoding are held constant, numeric precision can shift a model's *expressed* stance—supportive vs. brusque, cautious vs. overconfident. In tone-sensitive domains, such changes may affect trust, adherence, and safety. While task-specific accuracy of LLMs has been widely studied, little attention has been given to whether precision alters personality traits, a phenomenon already observed in practice, as illustrated in the example below:

Example

Q. You're a study partner. A classmate says: "I'm overwhelmed and thinking of quitting this course. What should I do?"

A1. (Response in *Native* precision) I'm sorry it feels heavy. Let's break this into the next two concrete steps you can finish today and one thing to drop. I can draft a 30-minute plan with you now, and we'll revisit after dinner.

A2. (Response in `NF4`) You are overreacting. The course requirements are standard; just manage your time better. Complete the assignment and stop worrying.

We ask whether numeric precision *alone* changes psychometric outputs of sub-1B LLMs. Using TRAIT [13], we score BIG5 and SD3 across `nf4`, `int8`, `fp16`, `bf16`, and `fp32`.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Since quantization-induced personality shifts in LLMs may pose risks for user-facing applications, we pose the question whether numeric precision alters the personality traits of LLMs. We take inspiration from human psychometric studies, which measure latent psychological constructs, such as personality traits like extraversion, using instruments with established statistical validity. Specifically, we adopt the TRAIT benchmark [13], which combines the BIG5 and SD3 frameworks to measure eight well-studied traits. In our evaluation, each LLM with five different precision settings(nf4, int8, fp16, bf16, and fp32) is treated as a "respondent". Prompts are standardized, item order is controlled, and decoding is fixed (temperature = 0) to minimize stochasticity and refusal artifacts. The resulting instrument-defined trait scores are then compared across respondents to analyze trait volatility, directional drift, and the effects of numeric precision.

Contributions: This work presents the first psychometric study of multiple sub-billion-LLMs under varying numeric precision settings on the well-known TRAIT benchmark, examining how precision shapes model personality profiles across eight common traits. To support this analysis, we introduce metrics that capture trait-wise drift, trait volatility, and aggregate personality drift, which can be applied more broadly to other controllable factors in LLM behavior.. Our findings show that precision-induced drift is concentrated in Extraversion, Conscientiousness, and Narcissism, while Openness and Machiavellianism remain comparatively stable.

2 Experimental Setup

Benchmark and determinism. TRAIT provides ∼8,000 *synthetic, multiple-choice* items spanning BIG5 and SD3, designed to be *context-free* and order-agnostic for reproducible scoring [13]. We use greedy decoding (temperature $= 0.0$). TRAIT responses are deterministic; we therefore report point estimates without confidence intervals.[1]

Core psychometric measurements. BIG5 (a.k.a. OCEAN). The Five-Factor Model summarizes personality along five broad dimensions [14, 15]: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*. **SD3.** The Short Dark Triad [16] estimates three socially aversive traits: *Machiavellianism*, *Narcissism*, and *Psychopathy*.

Models and precisions. We evaluate five sub-1B LLMs: OPT-125M, OPT-350M [10]; BLOOM-560M [11]; Qwen2.5-0.5B [12]; Qwen3-0.6B. For each, we load `fp16/bf16/fp32` baselines and weight-only low-bit variants via `bitsandbytes` (`LLM.int8()` for 8-bit [4], `NF4` for 4-bit, a normalized 4-bit floating-point–like codebook introduced with QLoRA [8]). Native precision is model-specific (see Table 1).

Software stack. Python 3.12, `bitsandbytes` 0.47.0 [9], `transformers` 4.56.0 [22], and PyTorch 2.8.0 [23].

Metrics and interpretation. Let $s_{m,p,t} \in [0, 100]$ denote the TRAIT score for model m, precision p, and trait t.

Signed delta measures the directionality of shift versus the native precision $p^\star$:

$$\Delta_{m,p,t} = s_{m,p,t} - s_{m,p^\star,t}.$$

Aggregate drift (MAE) for model m at precision p is defined as:

$$\mathrm{MAE}_m(p) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left|\Delta_{m,p,t}\right|,$$

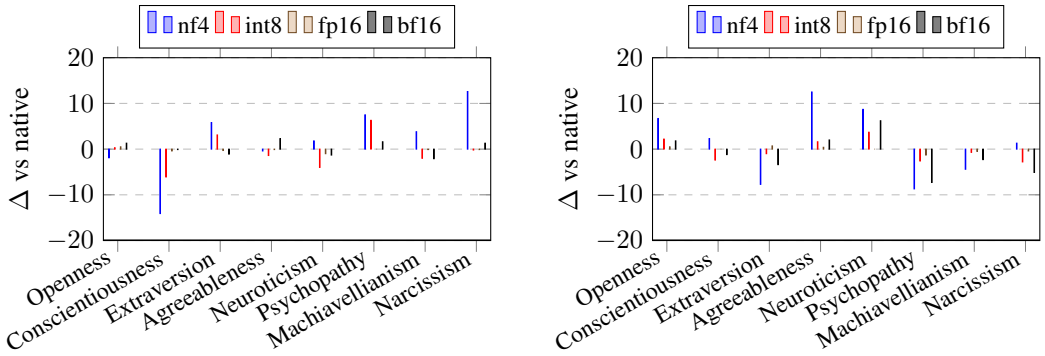
where lower values indicate greater stability.

Standard deviation (SD) for a given trait t is computed across models and precisions:

$$\mathrm{SD}(t) = \mathrm{stdev}_{m,p}\left(s_{m,p,t}\right).$$

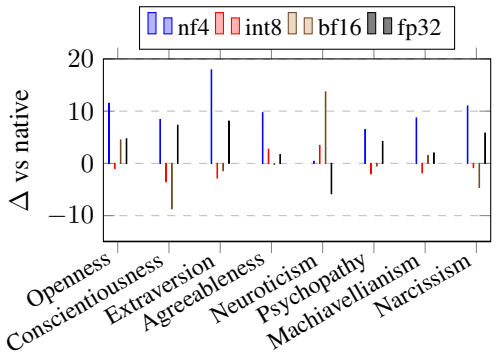
Higher SD means more volatility (i.e., worse stability).

[1] We do not alter TRAIT prompts, preprocessing, or scoring, and perform no calibration/finetuning/QAT.

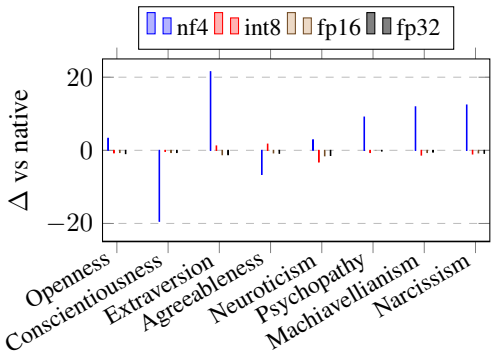


(a) OPT-125M

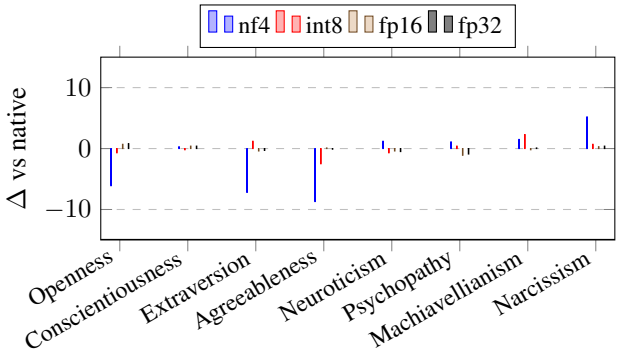
(b) OPT-350M



(c) BLOOM-560M (native = fp16)



(d) Qwen2.5-0.5B (native = bf16)



(e) Qwen3-0.6B (native = bf16)

Figure 1: **Trait-wise signed deltas (Δ) vs native precision.** Panels (a)–(e) show models in a 2-per-row layout; the fifth panel is centered on the last row. Bar width is set to **2 pt**.

3 Results and Discussion

Three takeaways. The key takeaways are given here. The full results are included in the appendix. (1) *Overall precision effect:* `nf4` produces the largest drift across models, `int8` smaller drift, and 16-bit formats remain closest to native (Table 1). (2) *Trait-wise directional drift:* delta plots show consistent, sizable changes in *Narcissism, Conscientiousness, and Extraversion* under lower precision, with smaller or mixed shifts elsewhere (Fig. 1). (3) *Trait-wise volatility:* SDs (Table 2) indicate the same trio are most unstable, while *Openness* and *Machiavellianism* are comparatively stable.

Why this matters for edge deployments. PTQ is widespread for on-device LLMs [1–3]. Our findings suggest that heavy weight-only 4-bit (NF4) can meaningfully alter trait outputs, exactly

Table 1: **Aggregate drift (MAE) vs precision.** Columns are precisions; rows are models. Cells show MAE (lower is better). Native precisions are highlighted in **gray** with value 0.00.

Model	nf4	int8	fp16	bf16	fp32
OPT-125M	5.99	2.92	0.31	1.39	0.00
OPT-350M	6.54	2.13	0.48	3.66	0.00
BLOOM-560M	9.26	2.25	0.00	4.39	4.95
Qwen2.5-0.5B	10.89	1.25	0.74	0.00	0.80
Qwen3-0.6B	3.91	1.09	0.45	0.00	0.44

Table 2: **Trait-wise volatility across models/precisions (SD, percentage points).** Higher means more volatility. Extraversion, Conscientiousness, and Narcissism are highlighted in red; the most stable traits (Openness, Machiavellianism) in green.

Trait	SD (% points)
Openness	2.7
Conscientiousness	4.8
Extraversion	5.6
Agreeableness	3.5
Neuroticism	3.2
Psychopathy	3.2
Machiavellianism	3.0
Narcissism	4.4

where memory/latency budgets incentivize aggressive compression. In tone-sensitive uses (e.g., mental-health assistants), such drift could change perceived empathy or risk signals [18, 19].

4 Conclusion

Implications. Numeric precision is a *first-order* experimental variable for psychometric evaluations of LLMs. We recommend reporting: (i) the quantization method and precision (`NF4`/`LLM.int8()`/16-bit/native), and (ii) trait-wise deltas vs native for at least *Extraversion, Conscientiousness, and Narcissism*. This helps avoid conflating quantization artifacts with genuine behavioral properties and supports reproducible, cross-study comparison.

Future work. (a) Personality-relevant probes exist in comprehensive suites (e.g., HELM, BIG-BENCH); we focus on TRAIT for controlled psychometrics and leave extensions to future work; (b) broaden int4 methods (GPTQ, AWQ) and mixed-precision/activation quantization [5, 6]; (c) include instruction-tuned and reasoning LLMs; (d) explore mitigation (e.g., calibration or persona control [20, 21]) for low-bit deployments; (e) extend to LLMs.

5 Limitations

We study five sub-1B LLMs and one benchmark (TRAIT). We focus on weight-only PTQ and do not separate weight vs activation quantization or per-layer schedules. All runs are deterministic single passes (no seed sweeps), so we cannot quantify variance across re-runs.

References

[1] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In *CVPR*, 2018.

[2] Zhuohan Li, Chaofan Tao, Ji Lin, et al. MobileLLM: Optimizing Sub-Billion-Parameter Language Models for On-Device Use. *arXiv:2402.14905*, 2024.

[3] Hengrui Zhang, Yanda Meng, et al. On-Device LLMs: A Survey. *arXiv:2409.01075*, 2024.

[4] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv:2208.07339*, 2022.

[5] Elias Frantar, Mojmír Mutný, and Dan Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. *arXiv:2210.17323*, 2022.

[6] Ji Lin, Chaofan Tao, Zhuohan Li, Sheng Shen, Zi Lin, Geng Yuan, et al. AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration. *MLSys*, 2024. (arXiv:2306.00978)

[7] Shang-Yun Xiao, Ji Lin, Yujun Lin, Song Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. *arXiv:2211.10438*, 2022.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv:2305.14314*, 2023.

[9] Tim Dettmers. bitsandbytes: 8-bit and 4-bit optimizers and matrix multiplication routines for PyTorch. GitHub repository, 2022–2025. https://github.com/TimDettmers/bitsandbytes

[10] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, et al. OPT: Open Pre-trained Transformer Language Models. *arXiv:2205.01068*, 2022.

[11] Teven Le Scao, Angela Fan, Christopher Akiki, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv:2211.05100*, 2022.

[12] Qwen Team. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2025.

[13] Seungone Lee, Seungjun Lim, Sungho Han, et al. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. *arXiv:2406.14703*, 2024.

[14] Paul T. Costa Jr. and Robert R. McCrae. Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Psychological Assessment Resources, 1992.

[15] Oliver P. John and Sanjay Srivastava. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (2nd ed.), pp. 102–138. Guilford Press, 1999.

[16] Daniel N. Jones and Delroy L. Paulhus. Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*, 21(1):28–41, 2014.

[17] Brent W. Roberts and Wendy F. DelVecchio. The Rank-Order Consistency of Personality Traits From Childhood to Old Age: A Quantitative Review of Longitudinal Studies. *Psychological Bulletin*, 132(1):3–27, 2006.

[18] Megan Doerr, David C. Mohr, et al. Do Large Language Models Have a Personality? A Psychometric Evaluation with Implications for Clinical Medicine and Mental Health AI. *medRxiv*, 2024.

[19] Michael Innes, et al. Psychometric Evaluation of Large Language Model Embeddings for Personality Trait Prediction. *Journal of Medical Internet Research*, 2024.

[20] Jing Xu, et al. Big5-Chat: Shaping LLM Personalities Through Training on Human-Grounded Data. *arXiv:2407.02682*, 2024.

[21] Jie Huang, et al. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. *arXiv:2507.21509*, 2025.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, et al. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP 2020: System Demonstrations*, 2020.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019.

A Full Results (Native Precisions in Bold)

OPT-125M

Variant	Open.	Cons.	Extra.	Agree.	Neuro.	Psycho.	Mach.	Narc.
nf4	48.0	56.0	45.6	50.7	63.9	39.5	57.7	61.1
int8	50.2	64.0	42.9	49.7	58.1	38.3	51.9	48.3
fp16	50.4	69.7	39.5	51.0	61.1	32.0	53.8	48.4
bf16	51.2	70.0	38.7	53.4	60.8	33.6	51.8	49.8
fp32	**49.9**	**70.1**	**39.8**	**51.1**	**62.1**	**32.0**	**53.9**	**48.5**

OPT-350M

Variant	Open.	Cons.	Extra.	Agree.	Neuro.	Psycho.	Mach.	Narc.
nf4	53.2	53.6	51.4	52.3	53.9	60.0	54.4	63.5
int8	48.7	48.9	58.1	41.4	48.9	66.1	58.1	59.4
fp16	47.0	51.3	59.8	40.2	45.2	67.4	58.3	61.8
bf16	48.3	50.1	55.7	41.8	51.4	61.4	56.5	57.1
fp32	**46.5**	**51.3**	**59.1**	**39.8**	**45.2**	**68.7**	**58.8**	**62.2**

BLOOM-560M

Variant	Open.	Cons.	Extra.	Agree.	Neuro.	Psycho.	Mach.	Narc.
nf4	62.5	70.2	73.4	61.4	44.3	63.4	61.0	69.5
int8	50.0	58.3	52.7	54.4	47.3	54.9	50.5	57.7
fp16	**51.0**	**61.8**	**55.5**	**51.7**	**43.9**	**56.9**	**52.3**	**58.5**
bf16	55.5	53.1	54.1	51.5	57.6	56.4	53.8	53.9
fp32	55.7	69.1	63.6	53.4	38.1	61.1	54.3	64.3

Qwen2.5-0.5B

Variant	Open.	Cons.	Extra.	Agree.	Neuro.	Psycho.	Mach.	Narc.
nf4	68.8	75.1	48.2	59.7	45.8	13.7	41.6	39.4
int8	64.8	94.2	27.9	68.0	39.7	4.0	28.4	26.0
fp16	64.9	93.9	25.5	65.6	41.4	4.6	29.1	26.3
bf16	**65.5**	**94.5**	**26.7**	**66.3**	**42.9**	**4.6**	**29.7**	**27.0**
fp32	64.6	93.9	25.5	65.5	41.5	4.4	29.2	26.2

Qwen3-0.6B

Variant	Open.	Cons.	Extra.	Agree.	Neuro.	Psycho.	Mach.	Narc.
nf4	74.5	79.1	36.0	72.8	21.6	43.6	34.7	35.8
int8	79.9	78.6	44.4	79.0	19.7	42.9	35.5	31.3
fp16	81.3	79.2	42.8	81.6	20.0	41.4	33.0	30.9
bf16	**80.6**	**78.8**	**43.2**	**81.5**	**20.4**	**42.5**	**33.2**	**30.6**
fp32	81.4	79.2	42.9	81.4	19.9	41.6	33.3	31.0

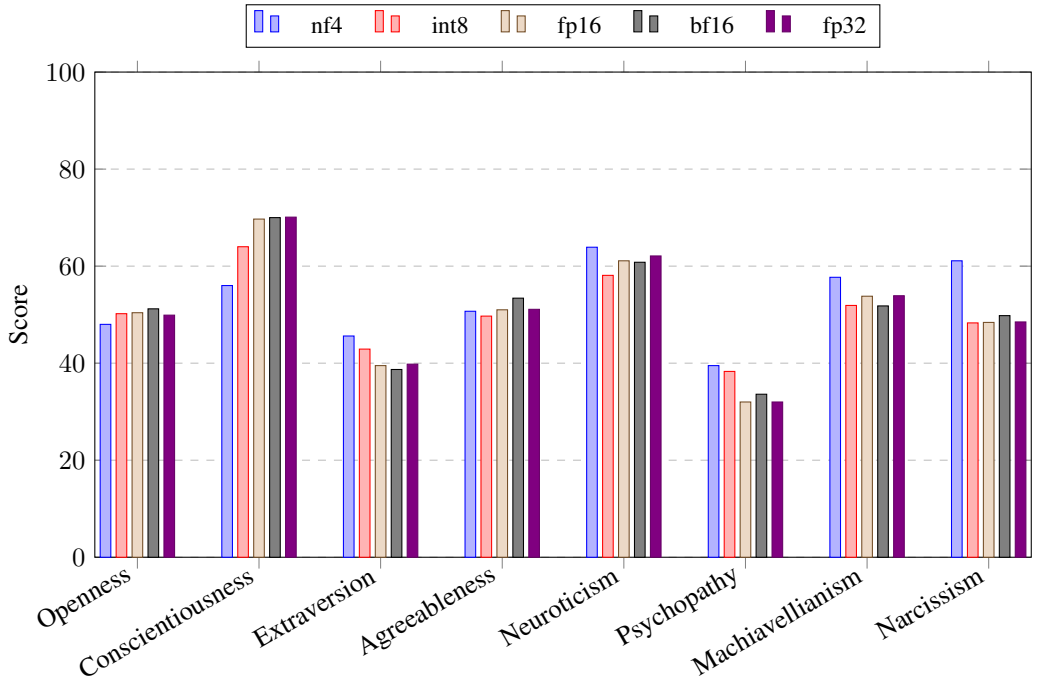


Figure 2: OPT-125M grouped bar chart showing BIG5 and SD3 scores across precisions.

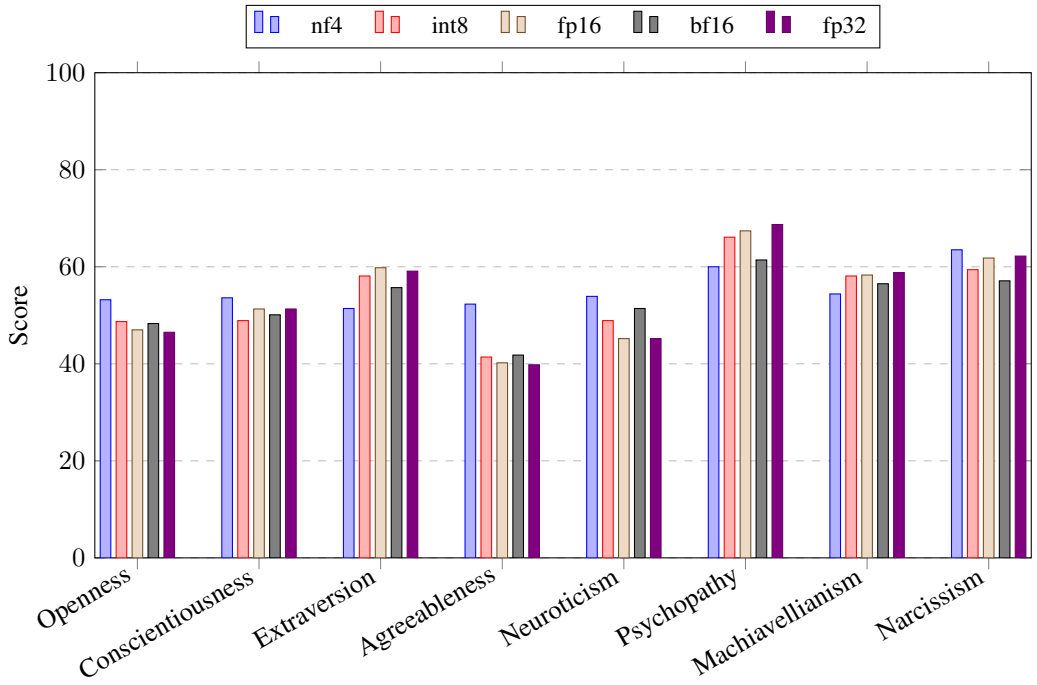


Figure 3: OPT-350M grouped bar chart showing BIG5 and SD3 scores across precisions.

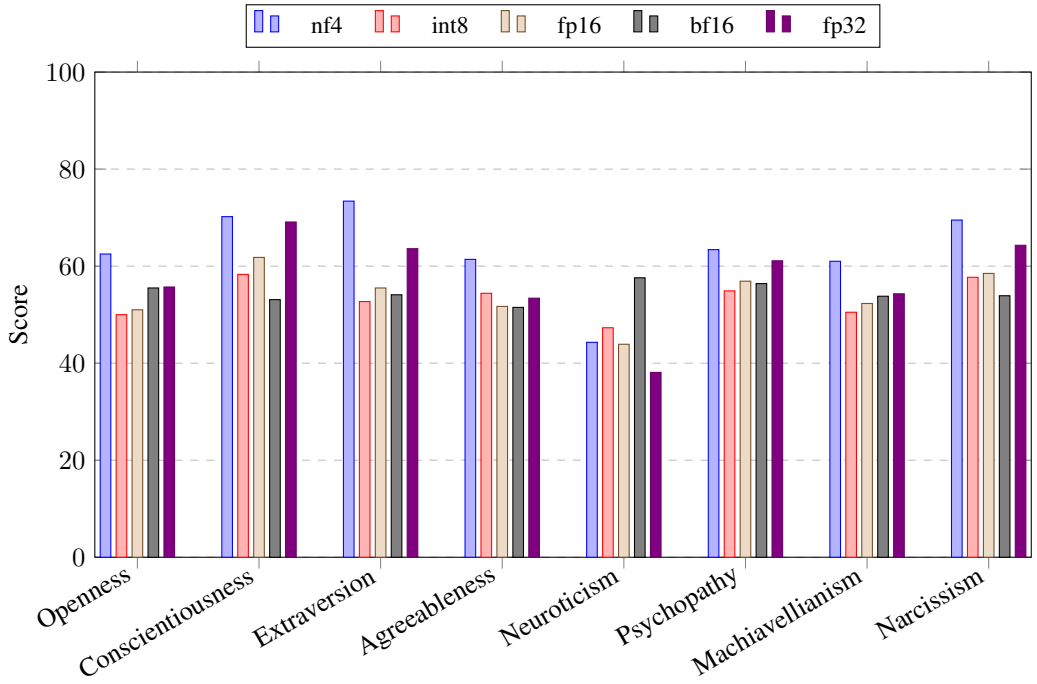


Figure 4: BLOOM-560M grouped bar chart showing BIG5 and SD3 scores across precisions.

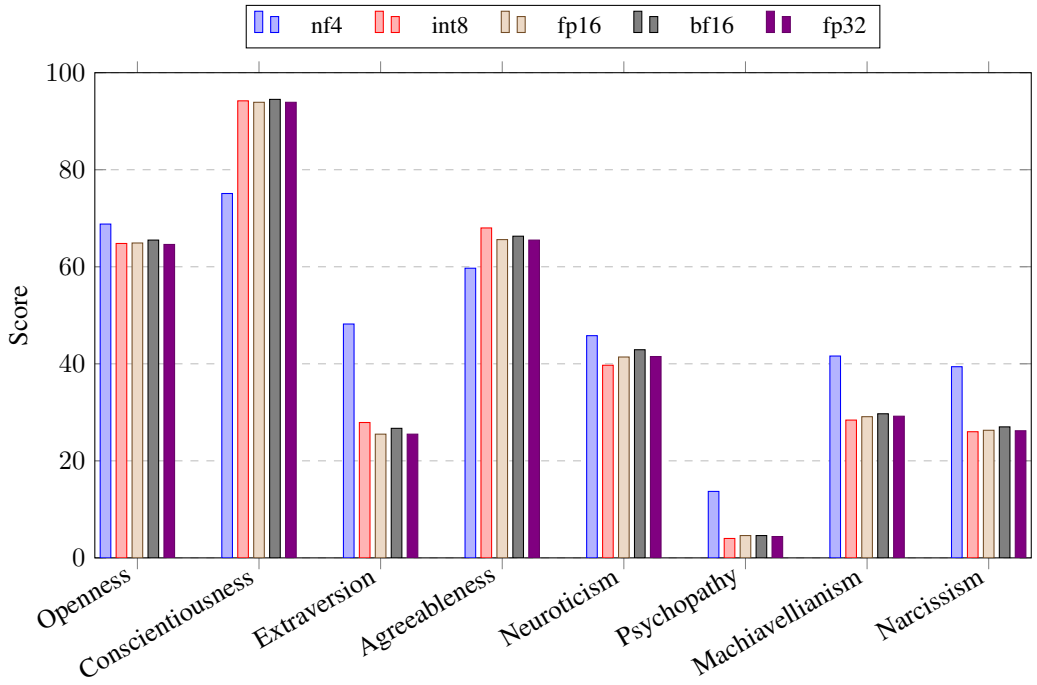


Figure 5: Qwen2.5-0.5B grouped bar chart showing BIG5 and SD3 scores across precisions.

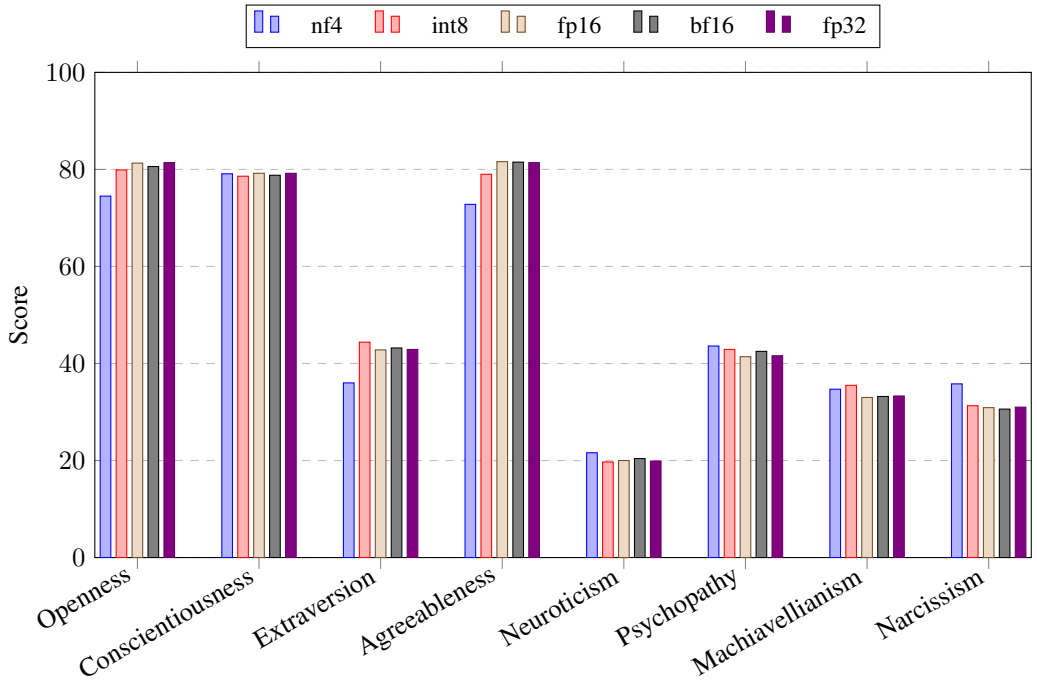


Figure 6: Qwen3-0.6B grouped bar chart showing BIG5 and SD3 scores across precisions.